ISOLATION AND CHARACTERIZATION OF RESISTANCE GENE ANALOGS (RGAs) IN SORGHUM

A Dissertation

by

JAE-MIN CHO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2005

Major Subject: Plant Pathology

ISOLATION AND CHARACTERIZATION OF RESISTANCE GENE ANALOGS (RGAs) IN SORGHUM

A Dissertation

by

JAE-MIN CHO

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Clint W. Magill (Chair of Committee)	Daniel J. Ebbole (Member)
Robert D. Stipanovic (Member)	Alan E. Pepper (Member)
Dennis C. Gross (Head of Department)	

May 2005

Major Subject: Plant Pathology

ABSTRACT

Isolation and Characterization of Resistance Gene Analogs (RGAs) in Sorghum.

(May 2005)

Jae-Min Cho, B.S.; M.S., Kyungpook National University

Chair of Advisory Committee: Dr. Clint W. Magill

The largest group of plant disease resistance (R) genes that share similar structures contains a predicted nucleotide-binding site (NBS) domain. NBS domains of this class of R genes show highly conserved amino acid motifs, which makes it possible to isolate resistance gene analogs (RGAs) by PCR with degenerate primers and homology searches from public databases. Multiple combinations of degenerate primers were designed from three conserved motifs (one motif was used for a subgroup-specific primer design) in the NBS regions of R genes of various plants. All combinations of primer pairs were used to amplify genomic DNA from sorghum. TIR-specific primer combinations showed no PCR amplification in sorghum. Homology searches identified many NBS-encoding sequences among the expressed or genomic molecular database entries for sorghum. Motif analysis of the sorghum NBS sequences that were identified in this study revealed eight major conserved motifs plus two additional highly conserved motifs, but no TIR-specific motifs. Phylogenetic analysis of sorghum NBS sequences showed tree topology typical of NBS-LRR genes, including clustered nodes and long-branch lengths. Eleven distinct families of NBS sequences, representing a highly diverse sample, were isolated from *Sorghum bicolor*. With two exceptions, sorghum RGA

families appeared to be closely related in sequence to at least one R-gene cloned from other species. In addition, deduced amino acid sequences of sorghum RGAs showed strong sequence similarity to almost all known non-TIR (Toll/Interleukin 1 Receptor)-type R-genes. Mapping with sorghum RGA markers revealed one linkage group containing four out of ten randomly selected markers, suggesting non-random distribution of NBS sequences in the sorghum genome. Rice sequences homologous to sorghum NBS sequences were found from two-way BLAST searches. Some of them were shown to be orthologs, when determined by using phylogenetic approaches which combined five different evolution models and tree-building methods.

DEDICATION

I dedicate this dissertation to my wife, Hee-Jeong Yang and my son, Myung-Hyun Cho, whose love and patience has encouraged me during my struggle with this research.

ACKNOWLEDGMENTS

I would like to thank Dr. Clint W. Magill, not only for his patient mentoring in science, but also for his impressive kindness. I feel fortunate for having had the opportunity to work with such a nice person.

I would like to acknowledge the committee members: Dr. Daniel J. Ebbole, Dr. Robert D. Stipanovic, and Dr. Alan E. Pepper for their advice and time.

I would like to thank all the people in the Magill laboratory for their help and support. I would also acknowledge all graduate students, faculty members, and staff members in Department of Plant Pathology & Microbiology for their support during my academic years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

GENERAL REVIEW

Flor's work with flax and rust disease (Flor, 1971) led to the concept that plant resistance (R) genes are responsible for phenotypic resistance against pests and pathogens containing corresponding avirulence genes. This relationship is known commonly referred to as the 'gene-for-gene' model or interaction. In this interaction, the R gene products somehow (directly or indirectly) recognize pathogen Avr gene products to trigger defense responses. These are often characterized by a hypersensitive response, which involves the cell(s) death and the local accumulation of antimicrobial compounds (Hammond-Kosack and Jones, 1996; Van der Hoorn et al., 2002). The recent ability to clone and sequence R genes has provided significant insight into their structure. R genes encoding proteins containing an N-terminal nucleotide-binding site (NBS) and C-terminal leucine-rich repeats (LRRs), represent the largest class of R genes in plants (Table 1) (Hulbert et al., 2001). NBS-LRR R proteins have two distinct N-terminal domain structures: the first is characterized by the Toll/interleukin-1/receptor (TIR) domain homologous to the *Drosophila* Toll and mammalian interleukin-1 receptors, and the second is characterized by a coiled-coil (CC) structure. Several conserved amino acid motifs exist in these domains, and some of them are subclass-specific so that the subclass of NBS-LRR genes can be predicted based on these motifs (Hammond-Kosack

---

This dissertation follows the style and format of Plant Cell.

Table 1. Classes of Characterized R Genes[a]

Class/gene	Interaction (host/pathogens)	Predicted protein structure	Complex locus[b]	References
1 *L*	Flax/*Melampsora lini*	TIR-NBS-LRR	No	Lawrence et al., 1995
M	Flax/*Melampsora lini*	TIR-NBS-LRR	Yes	Anderson et al., 1997
N	Tobacco/TMV	TIR-NBS-LRR	Yes	Whitham et al., 1996
P	Flax/*Melampsora lini*	TIR-NBS-LRR	Yes	Dodds et al., 2001
RPP1	*Arabidopsis*/*Peronospora*	TIR-NBS-LRR	Yes	Botella et al., 1998
RPP5	*Arabidopsis*/*Peronospora*	TIR-NBS-LRR	Yes	Parker et al., 1997
RPS4	*Arabidopsis*/*Peudomonas*	TIR-NBS-LRR	No	Gassmann et al., 1999
Bs2	Pepper/*Xanthomonas*	NBS-LRR	Yes	Tai et al., 1999
Dm3	Lettuce/*Bremia*	NBS-LRR	Yes	Meyers et al., 1998
Gpa2/Rx1	Potato/*Globodera*	NBS-LRR	Yes	Van der Vossen et al., 2000
	Potato/PVX (*Rx1*)		Yes	Bendahmane et al., 1999
I2	Tomato/*Fusarium*	NBS-LRR	Yes	Ori et al., 1997 Simons et al., 1998
Mi	Tomato/*Meloidogyne*	NBS-LRR	Yes	Milligan et al., 1998
	/*Macrosiphum*	NBS-LRR	Yes	Rossi et al., 1998 Vos et al., 1998
Mla	Barley/*Blumeria*	NBS-LRR	Yes	Zhou et al., 2001
Pib	Rice/*Magnaporthe*	NBS-LRR	Yes	Wang et al., 1999
Pi-ta	Rice/*Magnaporthe*	NBS-LRR	Yes	Bryan et al., 2000
Prf	Tomato/*Pseudomonas*	NBS-LRR	Yes	Salmeron et al., 1996
Rp1	Maize/*Puccinia*	NBS-LRR	Yes	Collins et al., 1999
RPM1	*Arabidopsis*/*Peudomonas*	NBS-LRR	No	Grant et al., 1995
RPP8/HRT	*Arabidopsis*/*Peronospora*	NBS-LRR	Yes	McDowell et al., 1998
	Arabidopsis/TCV (*HRT*)			Cooley et al., 2000
RPP13	*Arabidopsis*/*Peronospora*	NBS-LRR	No	Bittner-Eddy et al., 2000
RPS2	*Arabidopsis*/*Peudomonas*	NBS-LRR	No	Bent et al., 1994 Mindrinos et al., 1994
RPS5	*Arabidopsis*/*Peudomonas*	NBS-LRR	No	Warren et al., 1998
Rx2	Potato/PVX	NBS-LRR	Yes	Bendahmane et al., 1999
Sw-5	Tomato/*Tospovirus*	NBS-LRR	Yes	Brommonschenkel et al., 2000
Xa1	Rice/*Xanthomonas*	NBS-LRR	No	Yoshimura et al., 1998
2 *Cf-2/5*	Tomato/*Cladosporium*	LRR-TM	Yes	Dixon et al., 1998
Cf-4/9	Tomato/*Cladosporium*	LRR-TM	Yes	Jones et al., 1994 Takken et al., 2000 Thomas et al., 1997
3 *Pto*	Tomato/*Pseudomonas*	Protein Kinase	Yes	Martin et al., 1993
4 *Xa21*	Rice/*Xanthomonas*	LRR-TM-Kinase	Yes	Song et al., 1995
5 *HS1^{pro-1}*	Beet/*Heterodera*	Unique[c]	No	Cai et al., 1997
6 *Rpw8*	*Arabidopsis*/*Erysiphe*	Unique	Yes	Xiao et al., 2001
7 *mlo*	Barley/*Blumeria*	Membrane Prot.[d]	No	Buschges et al., 1997
8 *Hm1*	Maize/*Cochliobolus*	Toxin reductase	No	Johal et al., 1992

NBS = nucleotide binding site. LRR = leucine-rich repeat. TIR = domain with homology to the *Toll* gene of *Drosophila*, and the *Interleukin-1* receptor of mammals. TM = transmembrane domain. Domains are listed as they appear in the proteins from N to C terminal end.

[a]This table is quoted from Hulbert et al. (2001) with slight modification.

[b]'Complex locus' indicates the gene belongs to a tightly linked family of highly homologous genes.

[c]The predicted *HS1^{pro-1}* protein was originally reported to have a LRR-TM (Ellis and Jones, 1998).

[d]Predicted 60-kDa protein is membrane anchored with at least 6 membrane spanning helices.

and Jones, 1997; Young, 2000). In addition to their structural characteristics, research on NBS-LRR genes sheds light on the global genome organization, sequence variability and evolutionary history of R genes.

In this chapter, we review characteristics of NBS-LRR genes in the order of the following topics: structural organization of NBS-LRR genes, genomic distribution of NBS-LRR genes, phylogeny of NBS-LRR genes and evolution of NBS-LRR genes.

Structural Organization of NBS-LRR Genes

Structural characteristics are described below in the order in which domains are positioned in the proteins, starting at the N terminus. Recent genome-wide analyses of NBS-LRR genes in Arabidopsis and rice (Meyers et al., 2003; Zhou et al., 2004) provide the basis for the information that follows.

The N-Terminal Domain

The N-terminal domain begins with the start codon and ends at the start of the NBS domain. In this region, TIR or CC domains have been found with highly variable linker sequences. Some NBS-LRR R proteins contain a large N-terminal domain called the Toll/interleukin-1/receptor (TIR) domain, which has some similarity to the cytoplasmic signaling domain of the *Drosophila* Toll protein, the mammalian interleukin receptor (IL-1R), and a family of mammalian Toll-like receptors, one of which participates in recognition and response to lipopolysaccharides (LPS). Toll, IL-1R, and the mammalian Toll homologs all contribute to the immune response so are also involved in host defense

against pathogens (Medzhitov et al., 1997; Yang et al., 1998). The presence of the TIR domain in several R plant proteins suggests a role for this domain in signaling but not in ligand binding (Ellis and Jones, 1998). However, TIRs may also be involved in pathogen recognition. In a study of the alleles of the flax rust resistance gene *L* locus, two alleles with differing specificities were found to possess changes only in their TIR domains (Ellis et al., 1999). Thus, it appears that the TIR domain plays a role in signal transduction or pathogen recognition. Other NBS-LRR R proteins possess a putative leucine zipper (LZ) or coiled-coil sequence between the N terminus and the NBS domains. LZs are well known for their roles in homo- and hetero-dimerization of eukaryotic transcription factors as well as facilitating interactions between proteins with other functions. The LZs of Arabidopsis resistance protein RPS2 and RPM1 have 4 and 6 contiguous heptad sequences, respectively, that match the consensus sequence (I/R)XDLXXX (Landschulz et al., 1988). It is proposed that this domain facilitates the formation of a coiled-coil structure to promote either dimerization or specific interactions with other proteins.

In Arabidopsis, most TIR-NBS-LRR (TNL) proteins contain N-terminal consensus Met-Ala-polyserine [MA(S)n] residues that may enhance gene expression and protein stability. MA(S)n residues have been found in a variety of highly expressed genes, and mutations in these sequences have been shown to reduce reporter protein stability in plants (Sawant et al., 2001). No sequences related to MA(S)n are present at the N terminus of CC-NBS-LRR (CNL) proteins. Several conserved motifs (motifs TIR-1, TIR-2, TIR-3 and TIR-4) in the TIR domain of the TNL proteins are organized in

order and include ~175 amino acids. The CC domain is found at the N terminus of almost all CNL proteins of Arabidopsis and spans ~175 amino acids N-terminal to the NBS. The predicted CC motif is positioned from 25 to 50 amino acids from the N terminus in most CNL proteins. Twenty distinct motifs were identified in the N-terminal domain from CNL proteins using the program Multiple Expectation Maximization for Motif Elicitation (MEME) (Bailey and Elkan, 1995), and three CC domain patterns (CNL-A, CNL-B and CNL-C/D) with shared MEME motifs matched clades identified in phylogenetic analyses.

In rice, five major R gene organization types [one CNL type (four CNL types found) and one non-CC (XNL) type shown in Figure 1A] were identified from many different patterns of motifs when compared with the three motif patterns seen in CNL genes of Arabidopsis. The patterns of motifs indicate that the CC and non-CC genes group into subdivisions. The QLRD motif identified by Bai et al. (2002) contributes to the difference between CC- and non-CC genes (Figure 1A). This motif is usually located ~150 amino acids upstream of the P-loop, a region of high synteny that creates the motif in CC genes. These motifs in CC- and non-CC genes share a consensus sequence, indicating a common origin and a similar function.

The NBS Domain

NBS domains are responsible for ATP or GTP binding and hydrolysis (Tameling et al., 2002). The presence of the NBS suggests possible activation of a kinase or a role as a G-protein in signal transduction (Hammond-Kosack and Jones, 1997). Plant NBS domains

A

Arabidopsis

TNL

DVFPSFRGEDVRKTFLSHLLKEF IGPELIQAIRESRIAIVVLSKNYASSSWCLDELVEIMKC LGQIVMPIFYGVDPSDVRKQ
TIR-1 TIR-2 TIR-3

WRKALTDVANIAGEHS
TIR-4

CNL (CNL-B)

MGGCFSVSLPCDC VSQFSQLLCV SYIHNLSKNLASLQKAMRMLKARQYDVIRRLETEEF

RLSQVQVWLTSVLIIQNQFNDLLRSNEVELQRLC CSKDLKLSYRYGKRVIMMLKEVESLSSQ

TPSRDFDGMVGMEAHMEKMQSLLCLD

Rice

CNL

AATGVMSSLLxKLSSLLxEEY GVRREIEFLKDELESMNAALE KxWLxELRELAYDAEDCIDEF

GGAGAxGFLRKAAQRLKTLRA RHRIAxEIKELKARVEEVSERRKRYKLDDxAASPST

XNL

ATGAMGSLLGKLGELLxDEYx KGVRGEIxFLKDELESMHAFL xxLDQLDPQVKLWMR QLRDLAYDAEDCLDEFxYxx

xRHRIAxQIQELKxRLEEVSERRxRYGLD

Figure 1. Architectures of Motifs at the N-terminus and in the NBS Domains. These results were revealed from recent genome-wide analysis of *Arabidopsis* and rice (Meyers et al., 2003; Zhou et al., 2004). Conserved motifs are boxed. Homologous motifs are underlined or dot-lined. All major motifs identified by MEME in the N-terminal region **(A)** and in the NBS domain **(B)** are listed.

B
Arabidopsis

TNL

NxTPSRDFDDLVGIEAHLEKMKSLLCLES VGIWGPPGIGKTTIARALF
Pre-p-loop P-loop

DYGMKLHLQEQFLSEILNQKDIKIxHLGV RLKDKKVLIVLDDVD QLDALAGETxWFGPGSRIIVTTEDK
RNBS-A Kin-2 RNBS-B

NHIYEVxFPSxEEALQIFCQYAFQQNSPP EVAxLAGGLPLGLKVL EDKDLFLHIACFFNG
RNBS-C GLPL RNBS-D

MHNLLQQLGREIV
MHDV

CNL

QPQQDRQREMRQTPSKBSESELVGLEQNVKKLVGYL VGIYGMGGVGKTTLARQIF
Pre-P-loop P-loop

VKxGFDIVIWVVVSQEFTLKKIQQDILEK KRFLLVLDDIW NGCKVLFTTRSEEVC KVECLTPEEAWELFQRKV
RNBS-A Kin-2 RNBS-B RNBS-C

EVAKKCGGLPLALKVI CFLYCALFPEDYEIxKEKLIDYWIAEGFI VKMHDVVREMALWIA
GLPL RNBS-D MHDV

Rice

LVGIDGPREELIKLL VLSIVGMGGLGKTTLAQxVYN
Pre-P-loop P-loop

FDC RAWVCVSQNFDVxKLLR KRYLLVLDDV GSRIIVTTRIExVAx YKLEPLSDDDSWxLF
RNBS-I Kin-2 RNBS-II RNBS-III

ILKKCGGLPLAIKTI xxLExIRPILSLSYDDLPxHL KQCFLYCSIFPEDYxIxRDxLIRLWIAEGFIxE GExYFNELINRSFIQ
GLPL RNBS-IV RNBS-V RNBS-VI

CRMHDLMHDLAxSVS
MHDV

Figure 1. Continued.

show sequence similarity to nematode CED-4 and mammalian Apaf-1, which have been implicated in protease-mediated apoptosis (Li et al., 1997). Apaf-1 has also been shown to form oligomers (Srinivasula et al., 1998), which may be relevant in the function of plant R-gene NBS domains.

Although the exact boundary of the NBS domain is not defined, the NBS domain is composed of about 300 amino acids immediately following the N-terminus region of NBS-LRR R proteins and NBS sequences. The NBS domain is characterized by several highly conserved amino-acid motifs with variable regions between motifs. Within the NBS domain itself are found eight conserved motifs – P-loop, RNBS-A, Kin-1a, RNBS-B, RNBS-C, 'GLPL' sites, RNBS-D and MHDV (Meyers et al., 1999) – that differ in pattern between TIR and non-TIR subgroups. Three kinase motifs (P-loop, Kin-2a, kin-3(RNBS-B)) and 'GLPL' are highly similar in both groups. RNBS-A and RNBS-D motifs are dissimilar. In the RNBS-A motif region, while most TNL proteins contain a stretch of conserved amino acids with the consensus sequence LQKKLLSKLL, non-TNL proteins typically contain a distinctive amino-acid motif (FDLxAWVCSQxF). The TNL RNBS-D motif (FLHIACFF) is different from the non-TNL RNBS-D consensus sequence (CFLYCALFP). Motif RNBS-C also shows low similarity between TIR and non-TIR grouped R genes (Figure 1B). A single residue in the highly conserved motif (LLVLDDVW/D) within the NBS known as kinase-2 can be used to predict the presence of the TIR domain with 95% accuracy: a tryptophan (W) residue is found in non-TIR proteins whereas an aspartic acid (D) residue is found in TIR-containing proteins. Overall, these motifs are so diagnostic that it has been possible to develop degenerate

primers that specifically amplify either one of the two groups of NBS-LRRs (Pan et al., 2000).

Genome-wide analysis of 149 NBS-LRR-encoding genes in Arabidopsis confirmed two major classes that encode either 55 CC-NBS-LRR (CNL) or 94 TIR-NBS-LRR (TNL) proteins. The eight major motifs differed in their divergence within and between CNL and TNL groups, and in the same pattern as was observed for plant R protein homologs in general (Meyers et al., 1999). Comparisons revealed that the GLPL motif in the NBS domain of many TNL proteins contain some variations in the consensus core GLPL, and the most common variations contained the consensus GNLPL or SGNPL (although this is not shown in Figure 1B), showing lack of contiguous GL residues within the core of the motif. This is critical to the design of degenerate oligonucleotide primers for the amplification of R genes. The eighth conserved major motif called MHDV was highly conserved in CNL proteins, with a minor variation (QHDV) present in one CNL subgroup. The MHDV motif is slightly different in the TNL proteins, but it is clearly present and also shows high conservation of Met and His. The MHDV motif did not exist in any of the proteins that lacked an LRR in Arabidopsis, nor was it present in the divergent NBS-LRR (NL) proteins. It is thought this motif represents the C-terminal end of the NBS, at least when LRRs are present.

Another genome-wide analysis of NBS genes in rice suggested that the structure of the NBS domain is very similar to that in CNL gene products in Arabidopsis. Almost all of the NBS domains in the NBS-LRR genes contain one GLPL motif with the only

exceptions being two non-CC (XNL) genes, which contain three repeats of the GLPL motif. The conserved motif MHDV was found at the end of the NBS domains, but there is a clear difference between rice and Arabidopsis: in rice, the consensus sequence is not MHDV but MHDL (Figure 1B) and the MHDV motif is more diverse in the rice genes compared to the highly conserved MH residues in Arabidopsis.

Introns in the NBS region could be more common in cereals than in dicots. In 20 characterized dicot NBS-LRR R genes, only members of the Arabidopsis *Rpp8/Hrt* gene family have introns in the NBS domain. However, three characterized cereal resistance genes have introns in their NBS region, that is, *Mla1* (Zhou et al., 2000), *Pi-ta* (Bryan et al., 2000), and *Pib* (Wang et al., 1999). The most common intron position in cereals is at the N-terminal side of the kinase-2 motif, as is true in *Pi-ta* gene. *Pi-b* has a single intron between the RNBS-B and GLPL motifs. Arabidopsis *Rpp8/Hrt* genes have two introns in the NBS region, but the first intron is located before the RNBS-B motif and the second one is 21 amino acids upstream of the glycine residue in the GLPL motif. No rice genes showed similar intron positions for the *Rpp8* gene. Introns in the NBS region do not exist in TNL proteins in Arabidopsis, while two types of intron positions are found in CNL proteins.

The LRR Region

Leucine-rich repeats (LRRs) consist of repeated imperfect amino-acid segments that fold into solvent exposed β-strand β-turn structures, and this domain is thought to be involved in ligand binding and pathogen recognition (Jones and Jones, 1997). LRR regions are

characterized by alternating patterns of conservation and hypervariability. The variability is highest for codons (x) positioned around the two conserved aliphatic amino acids in the LRR consensus xxLxLxx, and the number of LRR repeats varies among family members.

The precise start and number of LRRs has not been well defined in many NBS-LRR proteins. In genome-wide analysis of Arabidopsis LRR regions, there were ~65 amino acids between the NBS and LRR domains in TNL proteins. Meyers et al. (2003) designated this non-LRR region the NL linker (NBS-LRR linker). In CNL proteins a short conserved NL linker was identified at ~40 amino acids C-terminal to the NBS domain. The motif for this linker was conserved within the different CNL classes but varied among classes. Truncated version (TN and CN proteins) NBS-LRR genes show lack of the LRR (Meyers et al., 2002) and there is no evidence of the NL linker protein sequences. A conserved NL linker motif (EENFVTVLDGQ) identified in rice is similar to the linker sequence found in the predicted products of C/D types of CNL genes in Arabidopsis, but is not similar to the other linkers found in Arabidopsis.

The C-terminal boundary of the LRR region was defined as the point at which LRRs no longer could be recognized. In Arabidopsis, LRRs constitute approximately half of the C-terminal region in the TNL proteins and nearly the entire C-terminal region in CNL proteins. The average TNL LRR domain and CNL LRR domain contained a mean of 14 LRRs with ~10 distinct MEME motifs that spanned ≥350 amino acids. Duplication patterns were recognized clearly as repeated MEME motifs in several CNL and TNL LRR domains, suggesting that duplications of LRRs accounted for much of the

variation in length. A total of 25 different LRR motifs were identified in the rice proteins by MEME (Table 2). The number of LRR repeats in any one gene ranged from 3 to 40. The precise pattern of LRR repeats varied widely, while the basic pattern was conserved as LxxxLxxLxxLxxLxLxxC (or T, S)xx. The occurrence and distribution of LRR motifs among NBS-LRR genes is also quite different.

The C-Terminal Domain

The size and composition of sequences in the C-terminal domain of genes in the CNL group is markedly different from those of TNL proteins in Arabidopsis. The difference in the C-terminal domain accounts for much of the increased total length of TNL versus CNL proteins. The CNL proteins have subgroup-specific (CNL-A, CNL-B and CNL-C/D) conserved motifs present in the 40- to 80-amino acid C-terminal domain, whereas the TNL proteins have a large number of non-LRR conserved motifs spanning ~200 to 300 amino acids (approximately as large as the LRR domain). The putative nuclear localization signal (NLS) [Deslandes et al. (2002) identified in the C-terminal domain of the Arabidopsis TNL:WRKY resistance protein RRS1] was also found in the C-terminal domain of most TNL proteins, but the particular amino acids of the NLS sequences were not conserved among TNL proteins. This suggests that the proposed NLS in RRS1 is either spurious or a unique variant of the conserved C-terminal domain found in most TNL proteins.

Table 2. LRR Repeats and Motifs[a]

(Sub)group	Motif[b]	Consensus sequence[c]
Arabidopsis		
TNL	Motif 1 (LDL)	MDLSYSRNLKELPDLSNATNLERLDLSYCSSLVELPSSI
CNL	Motif 1 (LDL)	IGNLVHLRYLDLSYTGITHLPYGLGNLKKLIYLNL
TNL	Motif 4 (end)	LHWLDLKGCRKLVSLPQLPDSLQYLDAHGCESLETVACP
CNL	Motif 4 (end)	LHTITIWNCPKLKKLPDGICF
Rice	LRR16	EIPPKVRHLSIxTDx
	LRR13	XMDLSHVRSLTVFGxx
	LRR14	LxxLKxLRVLDLEGCxxL
	LRR6	LxxIGxLxHLRYLxLRGTxIx
	LRR3	LPESIGKLxHLQTLDLRGT
	LRR4	LPxSIGKLKKLRHLxLxxxx
	LRR8	XLPxGIGKLTSLQTLxxVxIxxxx
	LRR20	FxVKKEDGYEIxQLKDMNELRxLxLxxxx
	LRR10	EAKEAKLxxKxHLxxLSLxWSx
	LRR5	LxxLQPPSNLKELxIxGYxGxxFPSW
	LRR7	XxxxGxFPxLRxLxIxDCPKLRxLP
	LRR27	GxLSRLPxWISSLxNLTKLxLxxxxL
	LRR9	LPxLGxLPSLRxLxLxxxxxL
	LRR11	XAFPKLEELVLxDMPNLEEWS
	LRR22	LPxxLxxLxSLKRLxIxNCPSLxSLPELGLPxSLEELxIxxCxxL
	LRR12	LxFEEGAMPKLERLELxxxx
	LRR19	xxxGIEHLxSLKELxxxx

[a]This table refers to the results from genome-wide analyses conducted by Meyers et al. (2003) and Zhou et al. (2004).
[b]The number assigned to the LRR repeats is the number output by the MEME analysis, and the order in the column generally reflects the region of LRR distribution in a gene.
[c]Underlined residues indicate possible LRR consensus matches (Jones and Jones, 1997). x denotes a variable site.

Genomic Distribution of NBS-LRR Genes

More than 150 NBS-LRR genes exist in the genome of *A. thaliana*. Richly et al. (2002) have listed a total of 166 NBS-LRR sequences, including 33 truncated sequences. These NBS-LRR sequences occur as 51 singletons and 40 clusters in their chromosomal arrangement. More NBS-LRR genes have been detected by Meyers et al. (2003) through the use of extensive manual re-annotation of the genomic sequence of the same species. Meyers et al. (2003) have listed 149 NBS-LRR genes and 58 truncated genes; the 149 non-truncated genes are distributed as 40 singletons and 43 clusters. In *A. thaliana*, TIR--NBS-LRR genes outnumber CC-NBS-LRR genes by roughly two to one, indicating either a recent amplification of the former family or loss of the latter family of genes (Cannon et al., 2002; Richly et al., 2002; Meyers et al., 2003). Arabidopsis NBS-LRR gene loci are not evenly distributed in the genome. Superclusters exist on chromosome 1 and 5, whereas chromosomes 2 and 3 are relatively deficient in NBS-LRR genes (Richly et al., 2002; Meyers et al., 2003). The clusters are thought to be involved in both the generation and maintenance of R-gene diversity.

Similar to the situation in Arabidopsis, the chromosomal distribution of the NBS genes is significantly non-random in rice: chromosome 11 contains about one-quarter of the NBS genes. Five hundred thirty-five NBS-encoding sequences, including 480 non-TIR NBS-LRR genes, were identified in rice. TIR-NBS-LRR genes have not been identified in the rice genome. Two hundred sixty-three genes (51 %) resided in 44 gene clusters. Counting 40 doublets and 17 triplets, 394 genes fall into the "clustered" distribution class. In total, 125 NBS singletons were dispersed over all the chromosomes.

The ratio of singletons to the total number of NBS genes in the rice genome (24.1 %) was similar to that in Arabidopsis (26.8 %; Meyers et al., 2003).

Phylogeny of NBS-LRR Genes

The phylogeny of NBS-LRR sequences divides into two major groups – TIR-NBS-LRR and nonTIR-NBS-LRR groups. Phylogenetic analyses performed by several groups with NBS-LRR R gene homologs collected from public molecular databases have consistently distinguished two clearly separated clades (Meyers et al., 1999; Pan et al., 2000; Cannon et al., 2002). TIR-NBS-LRR genes have not yet been identified and are probably absent in grass species, while nonTIR-NBS-LRR sequences are very common in these species (Pan et al., 2000). Previous efforts to isolate TIR-NBS-LRR sequences from grass species using TIR-specific degenerate primers or searching molecular databases uniformly failed. This is supported by the investigation of the whole genomes of Arabidopsis and rice. Recently, with the complete sequence of the genomes of *Arabidopsis thaliana* and rice, genome-wide analyses of the organization and evolution of NBS-LRR genes were carried out by several groups (Mondragon-Palomino et al., 2002; Richly et al., 2002; Baumgarten et al., 2003; Meyers et al., 2003; Zhou et al., 2004). Analysis of the *japonica* rice genome detected no TIR-NBS-LRR genes in the rice genome (Zhou et al., 2004). Although TIR domains are present in the rice genome, they are not associated with NBS-LRR genes. In *A. thaliana*, the phylogenetic analysis of Richly et al. (2002) and Meyers et al. (2003) have distinguished nine (seven TIR and two CC) and twelve (eight TIR and four CC) clearly distinguishable clades of NBS-LRR

genes, respectively. When considered with a report that TIR-containing NBS-LRR sequences are found in *Pinus* as well as in animals (Meyers et al., 1999; Pan et al., 2000), a model in which the common ancestor of Angiosperms and Gymnosperms contained both types of NBS-LRR sequences with the branch leading to modern grasses losing the TIR class of NBS-LRR sequences after divergence seems plausible.

Phylogenetic analysis reflects on diversity within the NBS-LRR family. Phylogenies of NBS-LRR sequences are characterized by long-branch lengths and closely clustered nodes, indicating ancient divergence into separate lineages followed by more recent diversification (Meyers et al., 1999; Pan et al., 2000; Cannon et al., 2002; Meyers et al., 2003; Zhou et al., 2004). The non-TIR branch of the NBS-LRR gene family is highly diverse (longer branch lengths than TIR branches), with several clades having originated prior to the split between Gymnosperms and Angiosperms (Cannon et al., 2002). Trees of non-TIR sequences are composed almost exclusively of species- or family-specific clades, though some branches containing sequences from multiple taxa do exist (Meyers et al., 1999; Pan et al., 2000). Within several of the major non-TIR clades, some well-sampled plant taxa are poorly represented or contain no resistance gene homologs (RGHs), suggesting either loss of particular RGH lineages in these taxa or growth or specialization in these RGH lineages in other taxa. This observation supports a birth and death model (this model interprets the expansion or contraction of gene clusters as the result of unequal crossover and the evolution of individual genes as the result of diversifying selection) of the evolution of this gene family (Michelmore and Meyers, 1998; Cannon et al., 2002). The TIR subfamily is more homogeneous,

suggesting either later divergence, more extensive structural constraints, or more concerted evolution than in the non-TIR subfamily (Cannon et al., 2002). The TIR group has relatively short branch lengths in contrast to the non-TIR group. Phylogenies of TIR-NBS-LRR sequences contain several distinct subgroups of sequences, reflecting recent diversification within individual species or closely related species. Some sequences are present multiple times within a single species (Cannon et al., 2002; Meyers et al., 2003). This indicates that some TIR-NBS-LRR sequences have diverged both prior to and since speciation. Nearly every branch of both TIR and non-TIR trees contains at least one confirmed R gene (Meyers et al., 1999, 2003), suggesting that most NBS-containing sequences are similar to known R genes and may therefore encode functional R proteins.

Evolution of NBS-LRR Genes

NBS-LRR genes are arranged as single genes and as clustered loci. The genomic analysis of Arabidopsis provides significant evolutionary information from the dissection of the phylogeny of NBS-LRR genes. Tandem gene duplications and duplication of individual or small groups of genes to unlinked loci (ectopic duplication) are, in general, the driving force for the distribution of NBS-LRR genes. The organization of NBS-LRR genes in arrays of members of the same clade is mainly a result of tandem duplications (Richly et al., 2002; Meyers et al., 2003).

New alleles are created by genetic recombination events between alleles or family members through re-assortment of the genetic variation created by mutation. Genetically linked gene families have more possibilities for recombination than simple

loci composed of single genes. Such crossovers can be intragenic or intergenic. Intragenic crossovers may generate novel alleles with different specificities (Ellis et al., 1999). Unequal crossover may change the number of family members in R gene clusters and rearrange them into new combinations (Parniske et al., 1997). In addition, the repeated action of equal and unequal recombination within a clustered gene family can homogenize them (known as concerted evolution) (Hickey et al., 1991; Walsh, 1987). The homogenizing effect of unequal recombination events slows divergence of family members and may actually hinder acquisition of new functions, such as the ability to recognize a novel class of *Avr* genes (Hulbert et al., 2001). In some R gene clusters, unequal recombination occurs frequently (e.g. in the *Rp1* and *Rp3* gene clusters of maize), whereas in others it is rare (e.g. in *Dm3* of lettuce and *Pto* of tomato) (Michelmore and Meyers, 1998). As a consequence, at loci similar to *Dm3* and *Pto*, orthologous genes from two different lines are more similar to each other than they are to paralogous genes within the same cluster.

Birth-Death models have also been proposed, emphasizing the importance of inter-allelic sequence exchange and diversifying selection (Michelmore and Meyers, 1998). The expansion or contraction of gene clusters result from unequal crossover and homogenization from gene conversions. In this model, divergent selection acting on arrays of solvent-exposed residues in the LRR results in evolution of individual R genes within a haplotype.

Recently, Baumgarten et al. (2003) have suggested that most of the genomic dispersion of NBS-LRR genes originates from duplication and translocation of entire

chromosomal segments (segmental duplication), rather than from small-scale ectopic duplication events. Most of the dynamic variation in NBS-LRR gene copy number occurs within local chromosomal regions. New NBS-LRR genes can arise and be lost through unequal crossing over, conversion, and an accumulation of mutations leading to either a pseudogene or a new function (Walsh, 1995; Michelmore and Meyers, 1998; Lynch and Force, 2000). Although accounting for a smaller fraction of gene duplication events, segmental duplication will have an impact on NBS-LRR gene family diversification. Segmental duplication could allow the preservation of many alleles that would not otherwise be maintained at a single NBS-LRR locus (Otto and Young, 2002). Tandem and segmental duplications distribute and separate NBS-LRR genes in the genome. It is, however, unclear by which mechanism(s) NBS-LRR genes from different clades are sampled into heterogenous clusters. Once physically removed from their closest relatives, the NBS-LRR genes might adopt and preserve new specificities because they are less prone to sequence homogenization.

Conclusions

Over the past few years, extensive genome sequencing (Arabidopsis and rice) and re-sequencing of R-gene clusters have provided valuable data, allowing much better understanding of the sequence organization, genome distribution and evolutionary history of plant R genes, especially NBS-LRR genes. The ancient nature of NBS-LRR sequences, their separation into distinct lineages and more recent diversification helps to explain the observed sequence diversity and structural features of this gene family. At a

genome level, extensive gene clusters are a striking property of most R genes that is probably related to a balance between creating new specificities and conserving old ones. The possibility of exchanges between clusters magnifies the opportunities for generating novel specificities. Future research must integrate our growing knowledge of R-gene sequence diversity and pathogen recognition and genome organization with parallel developments in new bioinformatics tools and coordinated efforts in structural and functional genomics. Building on the useful information mostly extracted from model plants, Arabidopsis and rice, information from non-model plants from a variety of plant species should lead to a much clearer understanding of the nature of resistance genes. Eventually it can be anticipated that with sufficient information, it may be possible to design effective resistance genes to interact with specific *avr* signals.

CHAPTER II

ISOLATION AND CHACTERIZATION OF RESISTANCE GENE ANALOGS (RGAs) IN SORGHUM

INTRODUCTION

A growing number of genes that confer resistance to a diverse spectrum of pathogens have been isolated from a wide range of plant species (Richter and Ronald 2000; Hulbert et al., 2001). These ″R″ genes have been classified into several groups based on the structural similarities of their predicted protein products. Most R proteins contain a nucleotide binding site (NBS) attached to a C-terminal leucine-rich repeat (LRR) of variable length. These domains participate in protein-protein interactions and signal transduction (Staskawicz et al., 1995). Such genes are called NBS-LRR R genes and represent the most prevalent class (Hulbert et al., 2001). So far, the only demonstrated role for NBS-LRR-encoding genes in plants is in disease or pest resistance (Michelmore, 2000).

NBS-LRR R genes are further subdivided into TIR- and non-TIR-groups based on the existence of a TIR domain at the N-terminal region. The TIR domain is named based on its original discovery from the *Drosophila* Toll protein and from mammalian interleukin-1 receptors and their homologs which are related to apoptosis of the cell in those organisms (Medzhitov et al., 1997; Yang et al., 1998). In plant R genes, TIR-homologous domains have been detected at the N-terminus region, suggesting function similar to Toll receptors in the resistance response in plants. This type of R genes has been categorized as TIR-NBS-LRR (TNL) genes. While TNL R genes contain a TIR

domain at their N-terminus, non-TIR R genes usually contain a coiled-coil (CC) domain instead of a TIR domain (Pan et al., 2000; Cannon et al., 2002). Most CC domains are leucine zippers. These two types of R genes somehow differentiate their signal pathways by two resistance signaling components EDS1 and NDR1: TIR-NBS-LRR proteins exclusively use EDS1, whereas NBS-LRR proteins with coiled-coil (CC) domains signal through NDR1 (Aarts et al., 1998).

The NBS domain is usually found in ATP- or GTP-binding proteins and is essential for the catalytic activity of these proteins since it functions directly in ATP- and GTP-binding (Saraste et al., 1990; Tameling et al., 2002). The NBS protein sequences can be assigned to separate subgroups based on the conserved motifs found within the larger domain (Traut 1994). The NBS domains of plant R genes can be categorized into two major types, which contain three distinguishing major motifs. These two types specifically match two subgroups of NBS-LRR R genes: one type is specific to TIR-NBS-LRR R genes and the second type matches non-TIR-NBS-LRR R genes. While TIR and non-TIR sequences have been isolated from dicot species, TIR-type genes have not been detected in genomic or expressed sequence tag (EST) sequences from any grass species (Meyers et al. 1999; Pan et al. 2000).

Many resistance gene analogous (RGA) sequences have been isolated from several groups of plant species using structural similarity within the NBS domain (Noir et. al., 2001; Madsen et al., 2003). NBS domains confer several advantages for identifying homologous sequences in new species – conserved motifs, a region of unvariable alignment, phylogenetic comparability and classification of NBS-LRR genes

by motifs within the NBS region (Meyers et al., 1999; Pan et al., 2000). In the NBS domain of plant R-genes, both TIR and non-TIR, a highly conserved backbone has been identified that is composed of eight major amino acid motifs. Some of these motifs are specific to the non-TIR class of proteins (Meyers et al., 1999, 2002, 2003; Zhou et al., 2004). The NBS-LRR sequences are so diverse that their overall homology is too low to be detectable by cross-hybridization. However, the existence of conserved motifs provides opportunities for the design of degenerate primers and the isolation of disease-resistance gene analogs (RGAs) by PCR from plant genomes. This approach has been successfully applied to isolate NBS-LRR genes from several monocot and dicot species (Kanazin et al., 1996; Yu et al., 1996; Leister et al., 1998; Shen et al., 1998; Noir et al., 2001; Madsen et al., 2003).

Genomic architecture of RGA sequences and actual R genes has been considered a source of diversity of the sequences and their evolution. Most NBS-LRR sequences are clustered in their chromosomal distribution. NBS-LRR sequences also reside in a certain chromosome more frequently than the other chromosomes. In Arabidopsis, 73.2 % of NBS-LRR genes (109 of 149) were distributed in 43 clusters (a cluster was defined as two or more NBS-LRR genes within a maximum of eight ORFs) (Richly et al., 2002). The largest cluster consisting of only NBS-LRR-encoding genes contained *RPP4*/*RPP5* plus seven NBS-LRR genes over a stretch of 90 kilobases (kb) on chromosome IV. Clusters contained combinations of TIR-NBS-LRR (TNL) or nonTIR-NBS-LRR (mostly CNL) genes with NBS-LRR related sequences TX-, TN-, or CN-encoding genes. The phylogenetic analysis revealed that the genes in clusters showed both monophyletic

and mixed patterns (Meyers et al., 2002, 2003). In rice, 51% of NBS genes resided in 44 gene clusters when Zhou et al. (2004) adopted Holub's (2001) definition of a gene cluster, which is a region that contains four or more genes within 200 kb or less. This percentage increased to 76% when tightly linked doublets and triplets were included in the estimation, which is similar to the distribution in Arabidopsis. About one quarter of the total number of NBS genes were placed in chromosome 11, showing a non-random chromosomal distribution pattern. The two largest clusters were both located on chromosome 11. The CC- and non-CC- (not TIR-relatives) types of genes similarly distributed on the chromosomes (Zhou et al., 2004). The physical structure of these clusters is thought to be involved in both the generation and maintenance of R-gene diversity.

Many studies of NBS-LRR sequences or resistance gene analogs demonstrated that R genes or NBS-LRR sequences in other plant species are also organized in large clusters. Clustering of resistance genes has been reported in maize for the *Rp1* (Collins et al., 1999), in barley for the *Mla* (Wei et al, 1999), in lettuce for a wide range of *Dm* loci (Meyers et al., 1998), and in flax for the *L* and *M* genes (Anderson et al., 1997; Ellis et al., 1999). The *M* locus of flax consists of 15 or more gene family members spread over a distance of less than 1 Mb (Anderson et al., 1997). The *Mla* resistance cluster of barley includes three NBS-LRR gene families within a 240 kb DNA interval on chromosome 5 (Wei et al., 1999). In the *Dm3* cluster of lettuce, at least 24 non-TIR NBS-LRR sequences were found to span approximately 3.5 Mb (Meyers et al., 1998). The clustered structure of NBS-LRR sequences has been shown from the characterization of resistance

gene analogs of a variety of plant species - Gymnosperms as well as Angiosperms – including soybean (Kanazin et al., 1996), coffee (Noir et al., 2001), maize (Quint et al., 2002), Medicago (Zhu et al., 2002), and barley (Madsen et al., 2003).

Sorghum is a member of the grass family and ranks fifth globally in value among the cereal crops (Doggett, 1988). Because sorghum is related to other cereals with a genome size between rice and maize (750 megabase pairs [Mbp]; Arumuganathan and Earle, 1991) and because it has great natural diversity (Dje et al., 2000), sorghum is often used for comparative analysis within the grass family. Numerous NBS sequences have been identified from the grass family members through three sources: known R genes, related NBS encoding genes in public databases, and sequences isolated by PCR using degenerate primers (Bai et al., 2002; Quint et al., 2002; Madsen et al., 2003). However, data from sorghum contributes a very small portion of those NBS sequences in the grass family and have so far received little attention.

Here I report isolation and characterization of RGA sequences from *Sorghum bicolor*. In the present study, a number of RGAs were obtained from *Sorghum bicolor*, using both degenerate primers based on conserved motifs of the NBS domain and public database searches. The sequence characterization and diversity analysis of these RGAs is reported as well as their relationships with the NBS sequences of known R-genes from other plant species. Moreover, we found rice orthologous sequences of the sorghum RGAs to provide an invaluable source of clarifying the function of uncharacterized genes. We also mapped the RGAs to find any linkage group. RGA maps will be helpful in isolating new R genes and searching for selectable markers for resistance.

MATERIALS AND METHODS

Plant Material and DNA Extraction

One elite line of *Sorghum bicolor* (BTx623) was used for PCR amplification of RGAs. It is one of the parents in the sorghum recombinant inbred line (RIL) mapping population that is described later in detail.

Total genomic DNA was extracted and purified from either frozen or fresh leaf tissue as described by Murray and Thompson (1980) and Saghai-Maroof et al. (1984) except that tissue samples were extracted in CTAB solution at twice the described concentration for 3-4 h at 65˚C with occasional gentle inversion. The detailed extraction steps are as follows: one gram of fresh leaf tissue was ground with liquid nitrogen in a pre-chilled mortar and pestle. The powdered leaf tissue was transferred to a 50ml conical tube containing extraction buffer composed of 100mM Tris (pH8.0), 0.7M NaCl, 10mM EDTA, 2% CTAB, and freshly added 2-Mercaptoethanol. The tube with sample was incubated at 60°C for 30-60min. After adding and thoroughly mixing 10ml of chloroform/octanol (24:1) the tube was then centrifuged at 5,125 X g for 10min at 4°C. The aqueous phase was transferred to a new 50ml conical tube, and 2/3 vol. of isopropanol was added and centrifuged to precipitate the DNA. The DNA was washed with 76% EtOH/10mM NH_4OAc for 20 min and recovered after centrifugation. The DNA was then dissolved in 1.5ml of TE buffer (10mM Tris, 1mM EDTA, pH8.0) and quantified by spectrophotometry (NanoDrop ND-1000, NanoDrop Technologies). The DNA was checked for restriction digestibility and PCR compatibility.

Primers and PCR Conditions

A large set of degenerate primers (Table 3) previously designed by Pan et al. (2000) based on conserved motifs in the aligned amino acid sequences derived from known NBS-LRR R-gene sequences and RGAs were used to amplify RGA sequences from sorghum genomic DNA. Four degenerate and one non-degenerate primers were designed to correspond to the P-loop motif in the sense direction, while eight degenerate plus one non-degenerate primers were made corresponding to the 'GLPL' motif and TIR- or non-TIR specific RNBS-D motifs in the anti-sense direction. In total, forty-five combinations of degenerate primers were used with genomic DNA of sorghum cultivar BTx623.

PCR amplification was performed in a 25μl reaction volume containing the following reagents: 125ng of genomic DNA, each degenerate primer at 1μM and 1X REDTaq Ready Mix (Sigma). GeneAmp® PCR System 9700 was used for the amplification. After denaturation of the DNA template at 94°C for 4 min, amplification consisted of 35 cycles of denaturation at 94°C for 45s, annealing at 45°C for 30s, and elongation at 72°C for 1 min. The last round of elongation was for 10min at 72°C to increase the fraction of products containing an A overhang.

Cloning and Sequencing of PCR Products

PCR products were checked on a 1% agarose gel, and directly cloned using a TOPO TA cloning® kit (Invitrogen). Clones were sequenced using the Applied Biosystems model 373 XL or 377 XL automated sequencers in the Gene Technologies Laboratory (GTL) at Texas A&M University. Each insert of the appropriate size was sequenced in both

Table 3. Degenerate Primers Used to Amplify Resistance Gene Analogs (RGAs)

Primer name	Group	Motifs	Oligo sequences (5' → 3')[a]
Forward			
H1145	-	GGVGKTT	GGI GGI RTI GGI AAI ACI AC
H2016[b]	-	GGVGKTT	GGT GGG GTT GGG AAG ACA ACG
H2017	-	GGSGKTT	GGI GGI WSI GGI AAR ACI AC
H2018	-	GGLGKTT	GGI GGI YTI GGI AAR ACI AC
H2019	-	GGMGKTT	GGI GGI ATI GGI AAA ACI AC
Backward			
H1146	Universal	GLPL	IAR IGY IAR IGG IAR ICC
H2021[b]	Universal	GLPLAL	CAA CGC TAG TGG CAA TCC
H2020	Universal	GL/FPL/FAL/V	CAA NGC CAA NGG CAA NCC
H2022	Universal	GL/FPL/FAL/V	CAG NGC NAG NGG NAG NCC
H2023	TIR	FLDIACF	RAA RCA IGC SAT RTC IAR RAA
H2026	TIR	FLHIACF	RAA RCA IGC DAT RTG IAR RAA
H2024	non-TIR	LKRCFLY	RTA IAG RAA RCA ISK YAG
H2025	non-TIR	FAYCSLF	RAA IAR ISW RCA RTA IGC RAA
H2027	non-TIR	YCALFPE	YTC IGG RAA IAR IGC RCA RTA

[a]I=inosine, R=A/G, W=A/C, Y=C/T, N=A/G/C/T, S=G/C. D=A/G/T, K=G/T
[b]Non-degenerate primers were used to compare PCR efficiency against degenerate primers.

directions using universal primers M13 and T7.

Database Searches for Sequences That Encode NBS Motifs Characteristic of R Proteins

BLAST version 2.0.3. (Altschul et al., 1997) was used to search the GenBank molecular databases and WU_BLAST version 2.0 (Washington University) was used on TIGR (The Institute for Genomic Research) *Sorghum bicolor* Gene Indices (SbGI) in March of both 2003 and 2004. TBLASTN searches were performed on dbEST, dbGSS(genome sequence survey, a database comprised of BAC end sequence tags) and dbNR(non-redundant) at NCBI GenBank, and on TIGR SbGI. Eighteen known NBS-LRR R-gene sequences were used to query the databases: *Gpa2*(AF195939), *I2C-1*(AF004878), *L6*(U27081), *M*(LUU73916), *Mi*(AF091048), *N*(A54810), *Pi-b*(AB013448), *Prf*(U65391), *Rp1-D*(AF107293), *RPM1*(AF122982), *RPP1*(AF098962), *RPP5*(AAB58295), *RPP8*(AAC83165), *RPS2*(U14158), *RPS4*(AJ243468), *RPS5*(AF074916), *Rx*(AJ011801), *Xa-1*(AB002266). Searches were conducted using the N-terminal NB-ARC domain sequences as defined by BLAST search with conserved domain database (RPS-BLAST) or Pfam database. The threshold expectation value was set to 0.0001, a value empirically determined to filter out most irrelevant hits. Other numerical options were left at default values except for the ″number of descriptions″ which was changed to maximum level to find all sequences hit by query sequences. Sequences were filtered to remove exact duplicates that resulted from searching multiple databases, and to combine overlapped sequences based on 'Sequencher' program results.

Motif Analysis

Multiple Expectation Maximization for Motif Elicitation or 'MEME' (Bailey and Elkan, 1995) was used to analyze conserved motif structures among NBS sequences. MEME discovers motifs by using a statistical algorithm called expectation maximization in unaligned sequences with no *a priori* assumptions about the sequences or their alignments. MEME reports a profile that describes a mathematical pattern in the conserved sequences. An individual profile describing amino acid frequencies is generated for each motif. Each position in the profile describes the probability of observing each amino acid at that position. Matches between the profile and individual sequences are scored by the program for each amino acid along the width of the profile. Multiple MEME analysis was performed with settings designed to identify 20, 30, 40 and 50 motifs; increasing the number of motifs simultaneously separates related motifs in different class sequences. The program MAST (Bailey and Gribskov, 1998) was used to assess correlations between MEME motifs in the distance matrix.

Alignment and Phylogenetic Analysis of Sequences

For the purpose of alignment, predicted protein sequences of sorghum NBS sequences plus 16 known R proteins were trimmed to generate four different datasets containing sequences that spanned four different motif regions: P-loop to Kin-2, Kin-2 to GLPL, GLPL to RNBS-D, and RNBS-D to MHDV. Sequences were then aligned using CLUSTAL W (Thompson et al., 1994) with default options. The alignment was inspected manually to make certain the conserved motifs aligned accurately, whereas the

more variable sequences between motifs contained minor ambiguous alignments. Phylogenetic analyses, including distance, parsimony, and bootstrap analyses, were performed using PHYLIP package version 3.6 alpha 3 (Felsenstein, University of Washington). Bootstraping provided an estimate of the confidence for each branch point (Felsenstein, 1985). The trees were rooted using a sequence from *Apaf*-1 as an outgroup, which is closely related to plant NBS-encoding R proteins.

Sorghum RIL Mapping Population

The population used in this study was used to construct the RFLP map of Peng et al. (1999). Sorghum microsatellites were subsequently placed on this map as detailed by Kong et al. (2000) and Bhattramakki et al. (2000). By appending about 2500 AFLP markers to this map, Menz et al. (2002) constructed a high-density genetic map containing 2,926 AFLP, RFLP and SSR markers. The population consisted of 137 F_{8-10} recombinant inbred lines (RILs) developed by Dr. K. F. Schertz (USDA-ARS) by single-seed descent from the cross between the elite inbred line BTx623 and IS3620C.

Detection of Restriction Fragment Length Polymorphism

Restriction fragment length polymorphism (RFLP) in the parents was examined prior to segregational analysis of RGA probes. Genomic DNAs (10µg per lane) digested with *Bam*HI, *Eco*RI, *Eco*RV, *Hind*III or *Xba*I were used for the detection of polymorphism between the two parental sorghum lines – BTx623 and IS3620C. Electrophoresis (Maniatis et al., 1982), blotting to Hybond N+ membranes (Reed and Mann, 1985) and

hybridization (Helentjaris et al., 1986) followed established protocols. Overnight hybridization was at 65˚C and blots were washed once in 2X SSC, 0.5% SDS, once in 1X SSC, 0.1% SDS, and twice or four times in 0.1X SSC, 0.1% SDS according to the strength of signal from the membrane. The washed membranes were placed into Image Plate (IP, Fuji Film Co.) cassettes at room temperature for 1~2 days to develop readable radioactive signal. Clones that revealed polymorphisms in survey blots were used for the analysis of the RIL population.

Mapping of RGA Sequences

Linkage analysis was conducted using Mapmaker version 2.0 on a Macintosh operating system. The 'ri-self' (recombinant inbred) setting was used, with any heterozygous genotypes for all codominant markers being considered missing data. Two-point linkage analysis with 'group' function with LOD 4 and recombination frequency of 0.40 were used to sort the loci onto linkage groups (LG). Multipoint linkage analysis of loci within LGs was subsequently performed. Using the 'compare' and the 'try' commands, the likely orders of RGA markers within LGs were determined and compared to assess the most likely orders. The Haldane mapping function was used to transform recombination frequency into cM (Haldane, 1919).

BAC Screening

The sorghum BAC libraries constructed by Tao et al. (1998) and Woo et al. (1994) were purchased from TAMU BAC Center. In total, 13,440 BAC clones (average insert size ≈

157 kb) are placed onto ten membrane filters and cover approximately three genome equivalents of sorghum.

The hybridization was performed by the following procedure recommended by TAMU BAC center: The filters containing sorghum BAC clones were pre-hybridized with hybridization buffer containing 0.5M sodium phosphate, 7 % (w/v) SDS, 1 % (w/v) BSA and 1mM EDTA for at least two hours at 65 °C in a rotary hybridization incubator. The probe (25 – 200 ng) was denatured by heating at 100 °C for approximately 5-10 minutes, and radio-labeled at 37°C for at least 30 minutes by using *Ready-to-go*® DNA labeling beads (Amersham). The radio-labeled probe was then denatured at 100 °C for 3 minutes and added into pre-hybridized sorghum BAC filters. The filters were incubated at 65 °C for at least twelve hours in an incubator. After hybridization, the filters were washed by gently shaking in a mixture of 0.5 X SSC prewarmed to 65 °C and 0.1 % (w/v) SDS at 65 °C. Washing was replicated for a total of three times, 15-20 minutes each wash. The washed filters were removed from the hybridization tube and wrapped wet in plastic wrap. The filters were then placed with Image Plate (IP®) to develop signals from probe-hybridized DNA.

Rice Ortholog Detection with Two-Way BLAST and Phylogenetic Methods

Both sorghum-rice and rice-sorghum BLAST searches at the nucleotide level were run against *Sorghum bicolor* Gene Indices (SbGI) and *Oryza sativa* Gene Indices (OsGI) at TIGR. The cutoff value for ortholog pairs was set at e-value 1e-5.

Exactly the same sorghum and rice sequence domains were used for phylogenetic

methods. Multiple alignments for tree calculation were constructed from each group of homologs by the program ClustalW (Thompson et al., 1994). After translating into the longest open reading frame (ORF), the sequences containing stop codons through most of the region were removed from the alignment before calculating the phylogenetic tree. Sequences >99% identical to any other sequence were also removed from alignment. Different phylogenetic analyses give different tree topology according to the method used and the model of evolution assumed. There is no overall consensus among biologists as to which phylogenetic method best reflects the evolution of proteins. Thus, instead of choosing one arbitrary method, several different evolution models and tree-building methods were used. The program PHYLIP was used for the analysis. The Jones-Taylor-Thornton (JTT) model (Jones et al., 1992), Dayhoff's PAM matrix (Dayhoff et al., 1979), and Kimura's formula (Kimura, 1983) in the neighbor-joining (NJ) tree, maximum parsimony (MP), and maximum likelihood (ML) with the JTT model were all used to build distance-based or character-based phylogenetic trees for ortholog detection. The list of orthologs was made on a consistency principle – the ortholog was marked only if the majority of five phylogenetic methods used supported a given pairing of orthologs.

RESULTS

PCR Amplification of RGAs with Degenerate Primers

Diverse sets of degenerate primers that had been successfully used in tomato, wheat or

coffee to amplify PCR products containing sequences homologous to known disease resistance genes (Pan et al., 2000; Noir et al., 2001) were used with sorghum. They were originally designed based on subgroup-specific conserved sequence alignments of R genes or R-gene homologs. The use of different combinations of forward and reverse primers permitted evaluation as to whether TIR-specific NBS sequences can be amplified from sorghum DNA. A total of forty-five different primer sets could be combined from five forward and nine backward primers in this study (Table 3).

PCR amplification resulted in a product that appeared to be a single band on a 1% agarose gel (Figure 2A). The PCR products were cloned and a total of 37 clones were sequenced, including at least 3 clones and up to 6 per each PCR product. Almost every PCR product showed heterogeneous sequences suggesting the involvement of a multigene family. In order to avoid sequencing numerous potentially identical clones from heterogeneous PCR products (Figure 2B), DNA from additional clones was digested using restriction enzymes with target sites revealed in the sequenced clone. Clones that gave identical insert sizes and restriction fragment patterns were considered to be products of a single amplification event. No additional clones were sequenced unless a different restriction fragment pattern was observed.

While no TIR-specific primer combinations (10 sets) produced any target PCR products, eight combinations of universal or non-TIR-specific degenerate primers successfully amplified products from genomic DNA of *Sorghum bicolor* (Table 4). This result was consistent with the previous work that failed to amplify TIR-specific PCR products from several grass family members (Pan et al., 2000). As expected from the

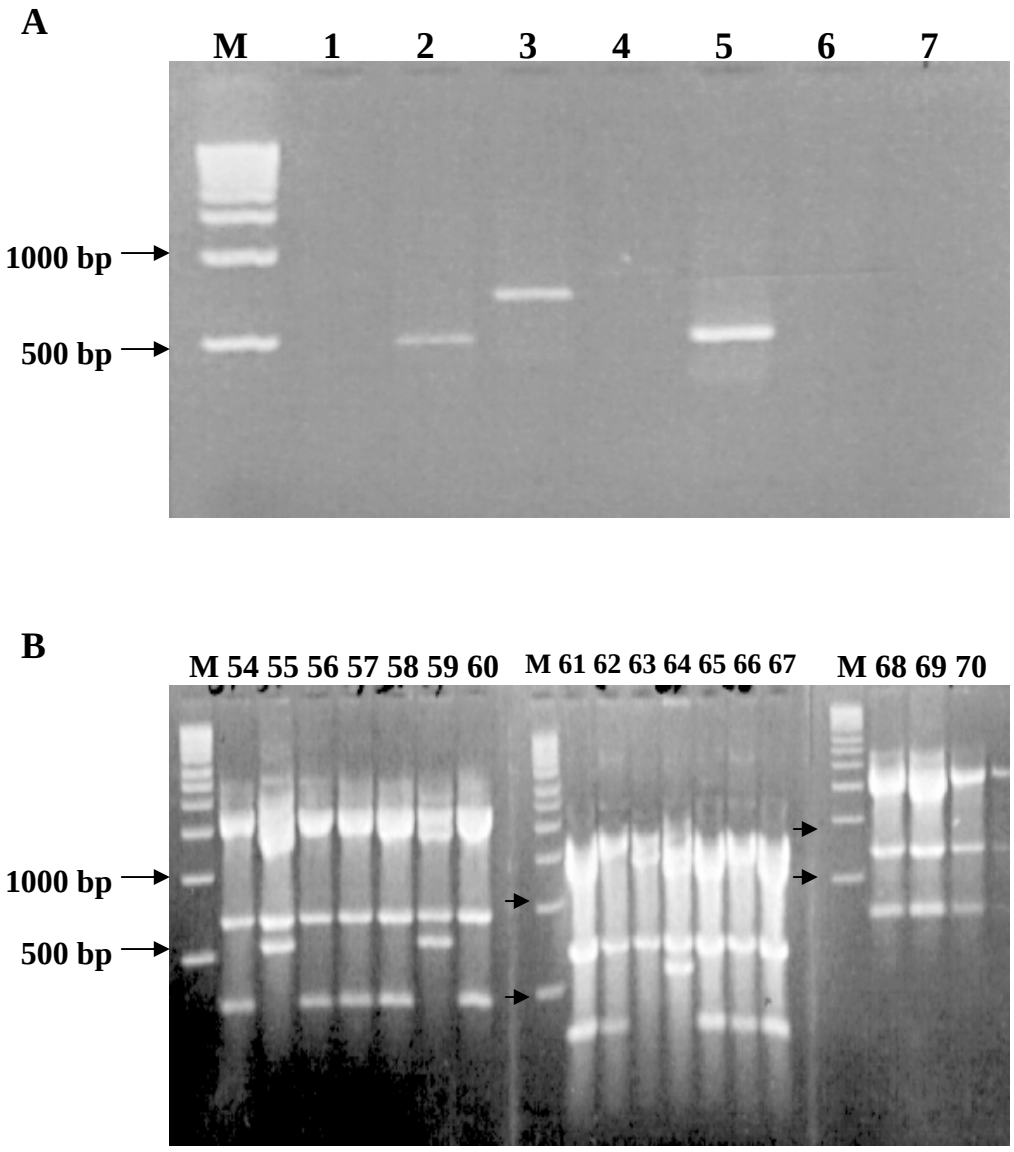


Figure 2. PCR Amplification of Resistance Gene Analogs (RGAs) **(A)** and the Presence of Heterogeneous PCR Product **(B)**. **(A)** Lane 1, TIR-specific backward primer; Lane 2, 3 and 7, non-TIR-specific backward primers; Lane 4, 5 and 6, universal backward primers. **(B)** Seventeen clones (RGA54 - 70) from single PCR product were digested with *Rsa*I. Two different fragment patterns were observed, suggesting that at least two different RGAs were amplified and cloned. Lane 63 contained vector with no insert.

Table 4. Characteristics of Resistance Gene Analogs (RGAs) Amplified from Sorghum

Primer pair	Detected PCR products[a]	Number of RGAs isolated[b]	Group[c]	RGA families represented[d]
F/B (TIR)				
all combinations	-	-		
F/B (non-TIR)				
H1145/H2027	+	2	non-TIR	-, H
H2018/H2027	+	1	non-TIR	G
H2019/H2027	+	2	non-TIR	G, G
F/B (universal)				
H1145/H2021	+	0		
H2017/H2021	+	1	non-TIR	H
H2018/H2021	+	1	non-TIR	B
H2019/H2021	+	0		
H1145/H1146	+	1	non-TIR	B
Total		8		

[a]PCR fragments of approximately 500 or 700bp in size as detected in 1% agarose gel fractionation.
[b]The number 0 indicates that the cloned fragment contained nonspecific sequence. The numbers 1 and higher indicate the number of clones that contain NBS-type sequence.
[c]Classification is based on the conserved sequence motifs shown in supplemental data in Appendix A1-10.
[d]RGA families were determined based on neighbor joining tree in figure 4. '-' indicates that the sequences were not included for phylogenetic analysis.

NBS domain size of known R-genes, major amplification products were about 500 or 700bp in size (Figure 2A). There is a possibility to amplify larger fragments than expected size because introns may exist in the NBS region. However, we only observed two band patterns unexpected in this study: one is smeared bands with unseparable PCR products and the other is the fragment of 600 bp in size. Although most PCR products with unexpected sizes were not considered for further analysis, those giving products near 600 bp were cloned and sequenced. When analyzed, these fragments were found to have no motifs characteristic of cloned R-genes. In total, eight unique sequences generated using conserved primers were identified as NBS-homologous sequences. For each of the eight amplified sequences, the deduced amino acids encoded all or parts of the internal motifs characteristic of the NBS-LRR R gene class (RNBS-A, Kin-2, RNBS-B and RNBS-C; Meyers et al. 1999). They all showed specificity to the non-TIR group of sequences based on the structures of RNBS-A and Kin-2 motifs (Appendix A-2 and A-3).

Database Searches

As seen in Table 5, as of March 2004 the NCBI nucleotide databases contained a total of 251,714 sorghum molecular sequences. Separate division entries included 161,813 Expressed Sequence Tags (ESTs), 89,534 Genome Survey Sequences (GSS) and 367 Non-Redundant (NR) sequences. Sorghum protein sequences (599 entries, as of March 2004) in the NCBI protein database were also used for this study. Although protein sequences could be found based on matching nucleotide sequences from the nucleotide

Table 5. Distribution of Sorghum Molecular Sequences

Resources	Type of sequence	Type of database	No. of entries	Release date
NCBI	Nucleotide	dbEST[a]	161,813	3/5/2004
		dbGSS[b]	89,534	
		dbNR	367	
		Total	251,714	
SbGI at TIGR	Nucleotide	TC[a, c]	18,659	12/22/2003
		Singleton ESTs[a, c]	18,409	
		Singleton ETs[a, c]	154	
		Total	37,232	
Univ. of Georgia	Nucleotide	EST library[a]	204,460	as of 3/12/2004
NCBI	Protein	dbProtein[d]	599	3/5/2004

[a]The EST sequences from three resources are mostly duplicated.
[b]dbGSS consists of BAC end sequences.
[c]*Sorghum bicolor* Gene Indices (SbGI) is composed of three types of expressed sequences: TC, singleton ESTs and singleton ETs. TC means tentative consensus sequences created by assembling ESTs into virtual transcripts. ET represents mature transcripts extracted and curated from sequences of GenBank.
[d]The protein sequences are usually matched to nucleotide sequences in dbNR.

databases, the protein database was preferred for identification of NBS sequences due to the advantage of no introns or codon degeneracy. The EST library (204,460 sequences in total) at the University of Georgia was also searched using internet-served database searching tools provided at that site. This search revealed new EST sequences not yet submitted to NCBI GenBank. The TIGR *Sorghum bicolor* Gene Index (SbGI) integrates research data from international *S. bicolor* EST sequencing and gene research projects. The TIGR Gene Index is designed to represent a non-redundant view of all *S. bicolor* genes, and contained a total of 37,232 unique sequences available for searches in this study.

The initial collection of NBS-LRR-like gene sequences was performed in March 2003 using a variety of predicted protein sequences from monocot and dicot NBS-LRR R genes. The trimmed NBS domain sequences of 18 NBS-LRR R genes were used as queries in BLAST searches. In total, 135 sequences including nine protein sequences were collected from all sources of molecular databases. Nine protein sequences were matched with 5 nucleotide entries (four BAC clones contained more than two protein entries). The respective nucleotide sequences spanning NBS domains were trimmed to compare with other nucleotide sequences. All the sequences were then analyzed with Sequencher v 3.0 program to remove duplicate sequences and the resulting number of sequences were narrowed to 70 unique sequences. Five unique sequences (the other three PCR products matched with three Database entries) were added from PCR products (Table 6). Another search using 12 seed alignment sequences of NB-ARC domains as defined by Pfam database was conducted in March 2004 against the same

Table 6. Summary of Sequences in Sorghum Molecular Databases Showing Homology to Known Plant R-Genes

Time searched	EST[a]	SbGI[a]	GSS[a]	NR[a]	PCR[b]	Total	Unique
March 2003[c]	42(2)[e]	37(8)	42(8)	13	8	134	75
March 2004[d]	19[f]	15	1	1		36	14
Total	61	52	43	14	8	178	89[g]

[a]EST, GSS, NR and SbGI are described in Materials and Methods.
[b]PCR=sequences isolated using degenerate primers to amplify R-gene homologs.
[c]NB-ARC domains of 18 plant R-genes (listed in materials and methods) were used as queries.
[d]NB-ARC domains of 12 seed alignment sequences defined by Pfam were used as queries.
[e]Numbers in parenthesis indicate sequences not detected in 2004 searches.
[f]Newly found sequences in 2004 were only added.
[g]The final number after removing duplicate or combining contiguous sequences.

databases. The increased number of sequence entries added fourteen new unique NBS sequences to the dataset. Finally eighty-four sorghum NBS sequences were identified from sorghum molecular databases (Table 6). All identified NBS sequences collected in this study are listed in Table 7.

In a 2004 search against the EST library at the University of Georgia, 58 sequences (0.028%) were obtained from 204,460 EST sequences (Table 8). This lower collection percentage of NBS sequences than that of NBS-LRR gene family members in the plant genome suggests that NBS-LRR genes were expressed at low levels during the various developmental stages from which mRNAs were extracted.

Motif Analysis

The 89 NBS sequences were translated and subjected to motif analysis. Except for 8 protein entries (AAD27570, AAM94294, AAM94295, AAM94297, AAM94396, AAO16686, AAO16692 and AAQ74890), most NBS sequences did not cover the whole region of the NBS domain, but spanned variable positions covering about one third of the NBS region. Because the sequences could not be aligned precisely, the program MEME (Multiple Expectation Maximization for Motif Elicitation) (Bailey and Elkan, 1995), which can be used with unaligned dataset sequences, was used in motif analysis.

Although 8 protein sequences also covered other regions (N-terminus and LRR region) of NBS-LRR genes, only the NBS region was used for all comparisons. Eight major motifs (P-loop, RNBS-A, Kin-2, RNBS-B, RNBS-C, GLPL, RNBS-D and MHDV; motif names starting from N-terminus of the NBS domain) have previously

Table 7. Sorghum NBS Sequences by Molecular Database Targeted for Searches

Database[a]	Identifier of NBS sequences[b]
dbEST at NCBI and Univ. of Georgia	AW285775, (AW286077), (AW286098), (AW286117), AW564339, AW672400, (AW925043), CD209645, CD211851, CD212839, (CD463246), (CF073050), (CF070823), (CF429173), (CF761005), (CF771727), BE355823, (BE359692), (BE594665), (BE595295), (BE595502), BE596218, (BE597203), (BE598046), (BE598072), (BE598263), (BE598264), (BE598785), BE599136, (BE599502), (BE600352), BG050233, (BG101746), BG412236, BG556059, BG557168, (BG948150), (BG948639), BI074536, (BI140073), (BI140459), (BI140694), (BI141181), (BI141270), (BI141271), (BI141394), (BI211333), BM317647, (BM322347), (BM322348), (BM322452), (BM323011), BM323307, BM324406, BM325057, (BM325821), BM325897, BM326535, BM327689, OX1_158_E07.b1, (OX1_158_E07.g1), RHOH_13_F05.g1, WS_10_C06.b1
dbGSS at NCBI	BH245455, (BH246001), BH246040, BH246056, BH246133, BH246154, (BH246155), BZ329687, (BZ330157), BZ330329, BZ331922, BZ334356, BZ337854, BZ338366, (BZ338367), BZ338669, BZ340437, (BZ341502), (BZ341503), BZ341506, BZ342222, (BZ348445), BZ343608, BZ345488, BZ346314, BZ348590, BZ349019, BZ349832, BZ350423, BZ350669, BZ367728, BZ369917, (BZ369918), (BZ422022), BZ423246, (BZ423379), BZ423689, BZ625990, BZ626449, BZ628476, BZ629156, (BZ629671), (BZ780807)
SbGI at TIGR	NP239121, NP239122, NP239123, NP239124, NP853482, TC75876, TC76169, TC76961, TC77858, TC79065, TC79359, TC79945, (TC80065), TC80519, TC80849, TC80927, TC81018, TC81885, TC83499, TC85900, TC86205, TC87218, TC89312, TC89319, TC90621, TC90798
dbProtein at NCBI	AAD27570, AAM94294, (AAM94295), AAM94297, AAM94306, (AAM94319), (AAO16686), (AAO16692), (AAQ74890)
dbNR at NCBI	(AF186640), (AF186641), (AF186642), (AF186643), (AF186644), (AF527807), (AF527808), (AF527809), (AY144442), (AY369028)

[a]The number of sorghum database sequences available in this study were described in Table 5.

[b]Some EST sequences were written by EST clone names because they had not been submitted to GenBank, NCBI as of March 2004. Names in parenthesis indicate duplicated sequences. Names underlined indicates that they had the same sequences as PCR products.

Table 8. Sorghum ESTs Related to the NBS of Plant R-Gene Products[a]

Sorghum bicolor EST library (Code)	ESTs related to NBS	Sequence Total[b]	
		3’	5’
Pathogen induced: incompatible (PI)	13	7488	6720
Pathogen induced: compatible (PIC)	11	6429	5663
Immature panicles (IP)	9	6624	6720
Light-grown seedlings (LG)	6	7675	7580
Heat-shocked seedlings (HS)	3	-	-
Dark-grown seedlings (DG)	2	8825	9981
Drought-stressed (WS)	2	7104	7104
Embryos (EM)	2	7296	7104
Iron-deficient seedlings (FE)	2	-	-
Oxidatively-stressed leaves and roots (OX)	2	-	-
Acid- and alkaline-treated roots (RHOH)	1	-	-
Drought-stressed after flowering (DSAF)	1	-	-
Drought-stressed before flowering (DSBF)	1	-	-
Ethylene-treated seedlings (ETH)	1	-	-
Ovaries (OV)	1	3344	3344
Phosphorus-deficient seedlings (PH)	1	-	-
ABA-treated seedlings (ABA)	0	-	-
Callus culture/ cell suspension (CCC)	0	-	-
Nitrogen-deficient seedlings (NIT)	0	-	-
Pollen (POL)	0	-	-
Total ESTs	58	204,460	

[a]The searches were performed in March, 2004.
[b]The total number of sequences as quoted from EST library at the University of Georgia. Total number of ESTs used for searches as listed in the results page of each search.

been identified in the NBS region of plant R genes, and several of these motifs demonstrated different patterns depending on whether they were present in the TNL or CNL groups (Van der Biezen and Jones, 1998; Meyers et al., 1999). MEME identified the eight major motifs with highly variable flanking sequences. The sequences of major motifs are shown in Table 9 and their alignments are shown in Appendix A-1 through A-8. The configuration of the motifs - kin-2, RNBS-A and RNBS-D – revealed no evidence for the existence of TNL sequences in sorghum. No aspartic acid residue (D) was detected at the end site of the kin-2 motif consensus sequence (LIVLDDVW) (Appendix A-3). The single residue (W/D) at this site can be used to predict the existence of a TIR domain at the N-terminal region preceding the NBS domain. That is, a tryptophan (W) residue has been found in CNL sequences whereas aspartic acid (D) is characteristic of TNL sequences (Young, 2000). The conserved sequences of RNBS-A and RNBS-D were similar to those of rice (RNBS-I and RNBS-V) or Arabidopsis CNL genes (Table 9).

A GLPL motif was found most often in the form of contiguous GL residues and some variations (40.5%) were observed in the L site of GL residues with 7-V substitutions, 5-S, 2-F, 2-I, and 1-Q. The consensus GNLPL or SGNPL, which are the most common variations of contiguous GL residues within the core of GLPL motif in Arabidopsis TNL proteins (Meyers et al., 2003), did not match any consensus core GLPL identified in this study (Appendix A-6).

The first residue of the eighth major motif, MHDV, was mostly V instead of M and the fourth V residue most often was replaced by an M residue (62.5%) (Appendix

Table 9. Major Motifs in Predicted Sorghum NBS Amino Acid Sequences

Motif[a]	Group[b]	Sequence[c]	Sources
P-loop	-	VSIVGFGGLGKTTLAQxVYND	Sorghum
P-loop	CNL	VLSIVGMGGLGKTTLAQxVYN	Rice
P-loop	CNL	VGIYGMGGVGKTTLARQIF	*Arabidopsis*
RNBS-A	-	FDCRAWVSVSQxFDVKKLLKEILEQLxKD	Sorghum
RNBS-I	CNL	FDCRAWVCVSQNFDVxKLLR	Rice
RNBS-A	CNL	VKxGFDIVIWVVVSQEFTLKKIQQDILEK	*Arabidopsis*
RNBS-A	TNL	DYGMKLHLQEQFLSEILNQKDIKIxHLGV	*Arabidopsis*
Kin-2	-	RYLIVLDDVWDxDVW	Sorghum
Kin-2	CNL	KRYLLVLDDV	Rice
Kin-2	CNL	KRFLLVLDDIW	*Arabidopsis*
Kin-2	TNL	RLKDKKVLIVLDDVD	*Arabidopsis*
RNBS-B	-	ALPxNxxGSRILVTTRIxxVA	Sorghum
RNBS-II	CNL	GSRIIVTTRIExVAx	Rice
RNBS-B	CNL	NGCKVLFTTRSEEVC	*Arabidopsis*
RNBS-C	-	VYELKPLSDxDSRELFxKRAF	Sorghum
RNBS-III	CNL	YKLEPLSDDDSWxLF	Rice
RNBS-C	CNL	KVECLTPEEAWELFQRKV	*Arabidopsis*
GLPL	-	ILKKCGGLPLAIVTIGSLLAS	Sorghum
GLPL	CNL	ILKKCGGLPLAIKTI	Rice
GLPL	CNL	EVAKKCGGLPLALKVI	*Arabidopsis*
RNBS-D	-	CFLYLSIFPEDYEIxRDRLIRRWIAEGFI	Sorghum
RNBS-V	CNL	KQCFLYCSIFPEDYxIxRDxLIRLWIAEGFIxE	Rice
RNBS-D	CNL	CFLYCALFPEDYEIxKEKLIDYWIAEGFI	*Arabidopsis*
RNBS-D	TNL	EDKDLFLHIACFFNG	*Arabidopsis*
MHDV	-	DEGRVKxCRVHDMVLDLICSKSREENFV	Sorghum
MHDV	CNL	CRMHDLMHDLAxSVS	Rice
MHDV	CNL	VKMHDVVREMALWIA	*Arabidopsis*

[a]Motifs are listed in the order that they occurred in the NBS domain. Sorghum motifs were named after *Arabidopsis* descriptions (Meyers et al., 1999, 2002, 2003).
[b]N- or C-terminal sequences of sorghum NBS sequences could not be determined due to lack of full-length sequences for analysis.
[c]Consensus amino acid sequence derived from MEME. Related motifs in the NBS of CNL and TNL proteins are aligned. The MEME output for the major motifs is available in the supplemental data in Appendix A1-10. x indicates a nonconserved residue.

A-8). Although small numbers (16 NBS sequences) of sequence were compared for the MHDV motif in sorghum, the consensus VHDM is clearly different from MHDL in rice and MHDV in Arabidopsis (Meyers et al., 2003; Zhou et al., 2004).

Two additional motifs in the NBS found in rice are called RNBS-IV and RNBS-VI (Zhou et al., 2004). These motifs reside between GLPL and MHDV motifs and are separated by RNBS-V (RNBS-D in Arabidopsis) motifs. In sorghum, two NBS motif sequences were detected with similar consensus sequences: ILSLSYNDLPSHLKT for RNBS-IV (xxLExIRPILSLSYDDLPxHL) and KGGKSLEELGESYFNELINRSLIQPV D for RNBS-VI (GExYFNELINRSFIQ) (Appendix A-9 and A-10). An additional motif (TKEEWxKVYNSIGSGLENNPD) located just following the GLPL motif was identified and spanned, together with a RNBS-IV motif, most of the region between GLPL and RNBS-D motifs (Appendix A-9). MEME detected the pre-P-loop motif as defined by 41 amino acids as the consensus sequence with only one ambiguous site (PTxVDPRLTALYLEASELVGIDKPRDELIDFLLDEDAADEA) (Appendix A-11).

MEME also identified several minor motifs besides the eleven motifs described above. MEME was used with default values for the optional parameters except that the 'number of motifs' possible was changed to 50. This allowed detection of as many conserved blocks as possible as well as the detection of highly conserved motifs within the NBS domain. MEME counted as a motif any sequence that was conserved by at least two NBS sequences. Thus, the minor conserved motifs were diverse and could be used to compare their distribution pattern and to further classify NBS sequences into several sub-groups. Sixty-five NBS sequences were classified into 11 groups where each group

had at least one unique motif shown in a different number and color box as compared to other groups of sequences (Figure 3). The remaining 24 sequences were not considered for the classification because they could fit into any groups due to their short length.

Phylogenetic Analysis of Sorghum NBS Sequences

Phylogenetic relationships between amino acid sequences deduced for the sorghum NBS sequences and known R-gene products were investigated. A variety of R-genes of the NBS-LRR class listed in the GenBank database were included in the analysis (*I2C-1* and *Prf* from tomato, *GPA2* from potato, *RPM1*, *RPS2*, *RPS5*, *RPP1* and *RPP8* from *Arabidopsis*, *RP1-D* from maize, *Pib* and *Xa1* from rice, *M* and *L6* from flax, and *N* from tobacco). Four phylogenetic trees were developed using four different datasets of NBS sequences because of difficulty in direct comparison with all collected sequence fragments which do not span the same region. Sorghum NBS sequences were grouped into four different datasets based on their motif coverage in the NBS domain. Group 1 contained sequences that spanned at least from the P-loop to the Kin-2 motif (Figure 4A). Groups 2, 3 and 4 included data that covered the regions of Kin-2 - GLPL, GLPL – RNBS-D and RNBS-D – MHDV, respectively (Figure 4B,C and D). The neighbor-joining phylogenetic trees (Saitou and Nei, 1987) constructed from the amino acid alignments of NBS domains of these R-genes and sorghum NBS sequences are shown in Figure 4. The sequence alignments for phylogenetic analysis are presented in Appendix B. The tree has long-branch lengths and closely clustered nodes, reflecting a high level of sequence divergence. The sorghum NBS sequences could be grouped into 11 families

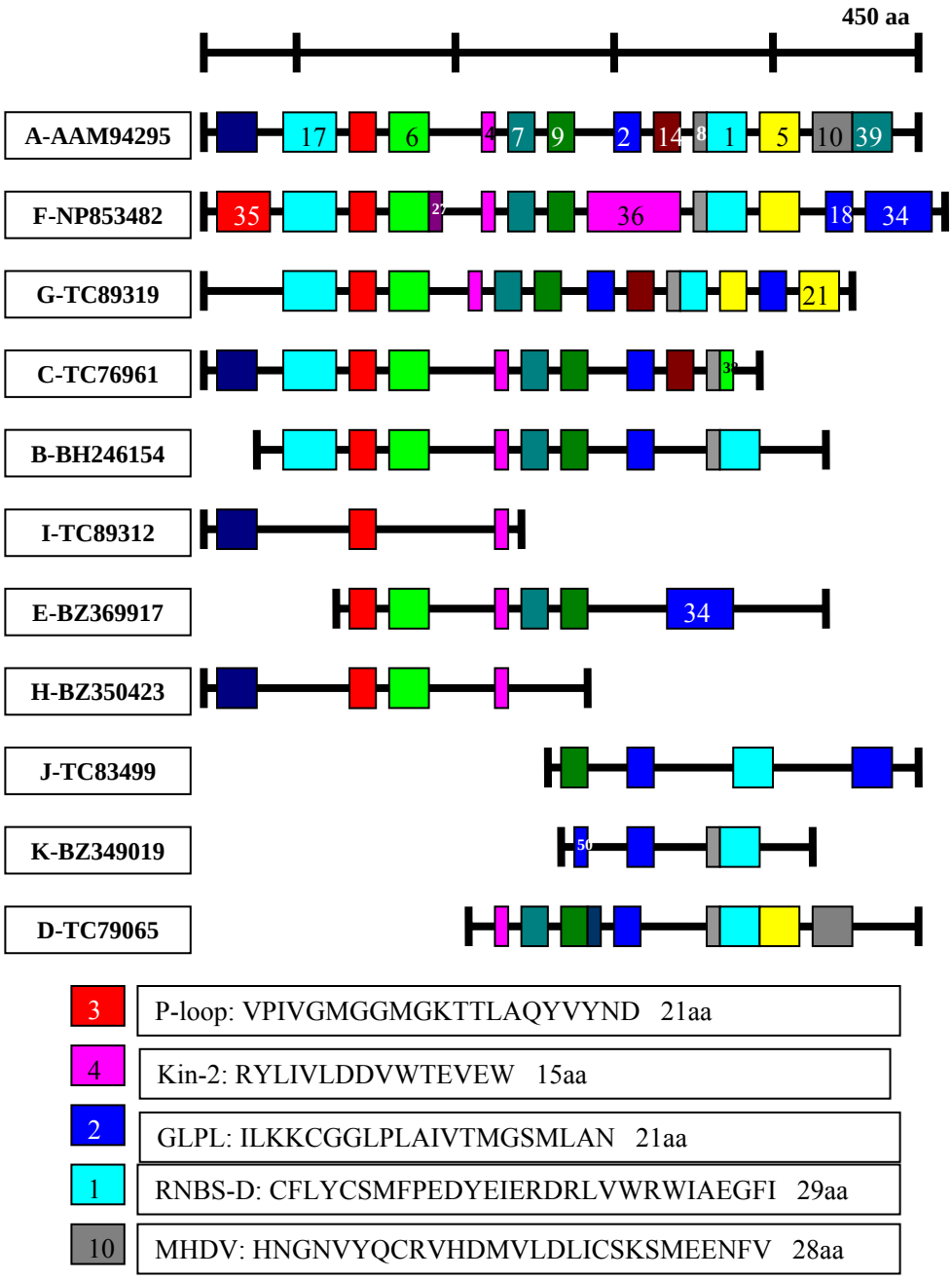


Figure 3. Motif Patterns in the NBS Domains of Sorghum NBS Sequences. Different colored boxes and numbers indicate separate and distinct motifs identified using MEME (Bailey and Elkan, 1995) and displayed by MAST (Bailey and Gribskov, 1998). The same colored boxes without numbers indicate the same motifs as shown in the top sequence. The consensus sequences of five major motifs are shown at the bottom of the figure.

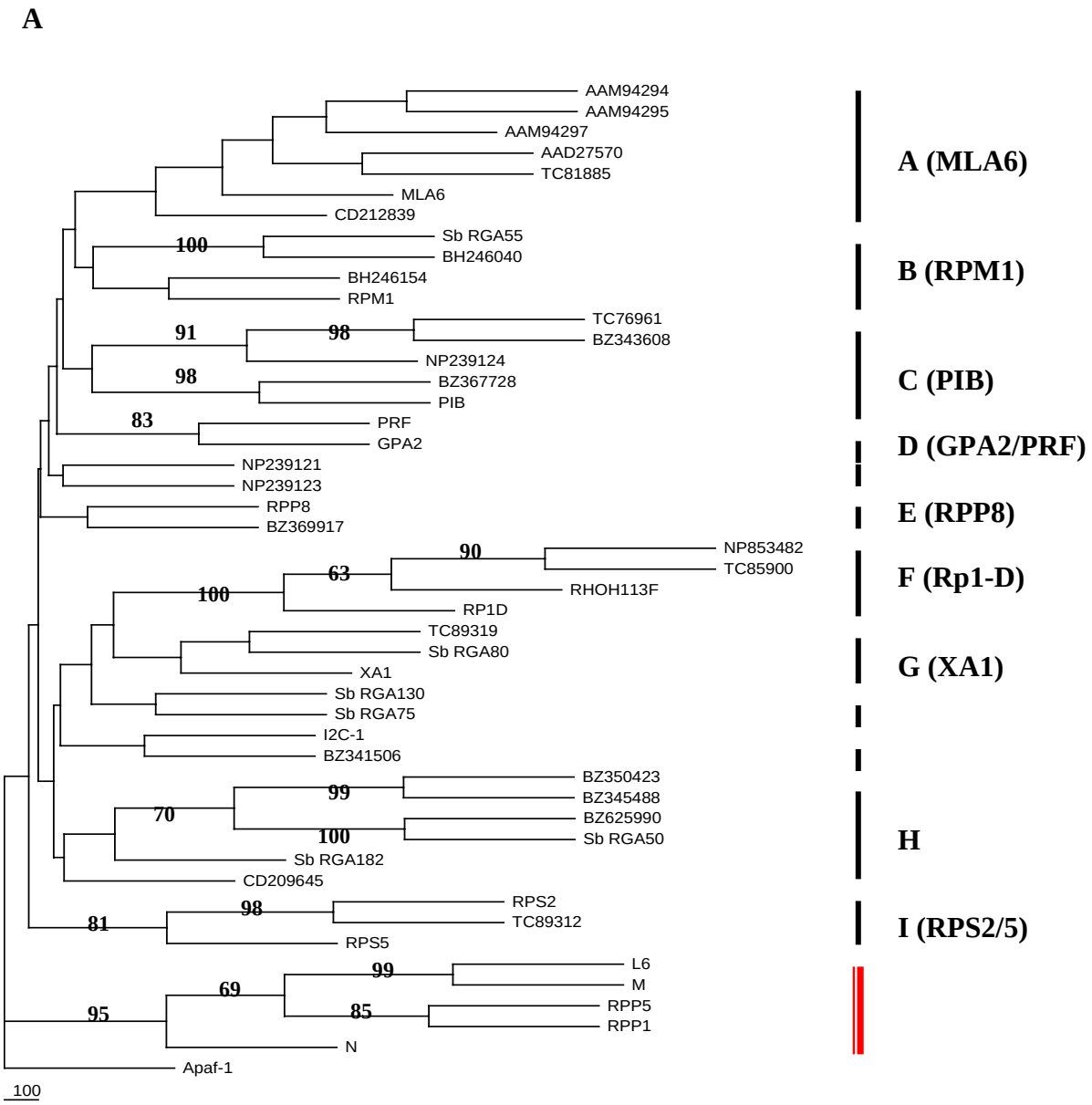


Figure 4. Neighbor-Joining Trees Based on Alignment of Amino Acids of Sorghum NBS Sequences and Cloned R Genes. Four different sets of sorghum NBS sequences were used for phylogenetic analysis: **(A)** P-loop to Kin-2, **(B)** Kin-2 to GLPL, **(C)** GLPL to RNBS-D, and **(D)** RNBS-D to MHDV. Bootstrap values are the percentage of 500 neighbor joining bootstrap replicates. Bootstrap values at or above 60% are shown. Bars on the right represent sorghum RGA families discussed in the text. Layer bars represent TIR-NBS-LRR R gene members. Unnamed bars indicates branches are not in agreement with those in other trees.

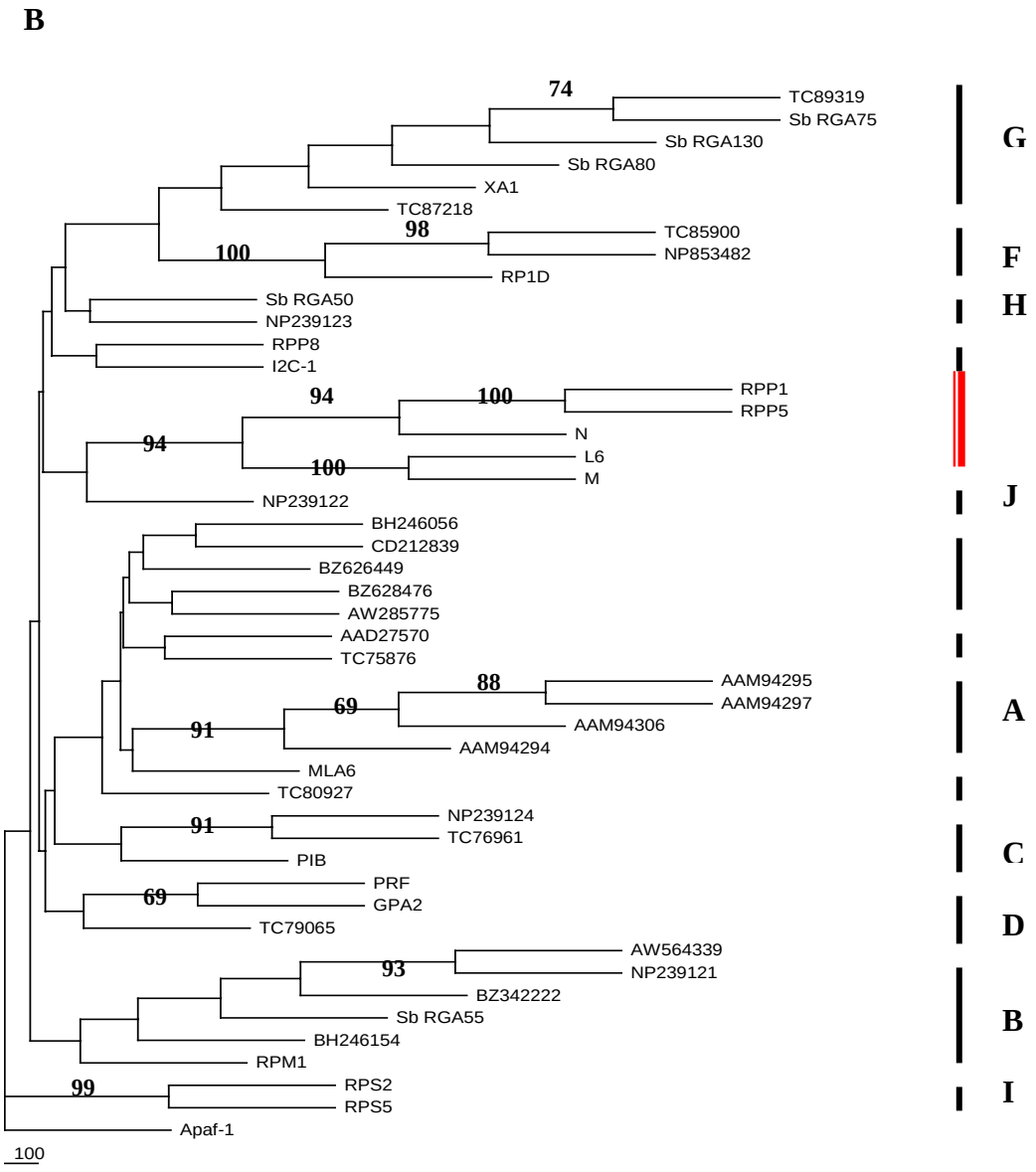


Figure 4. Continued.

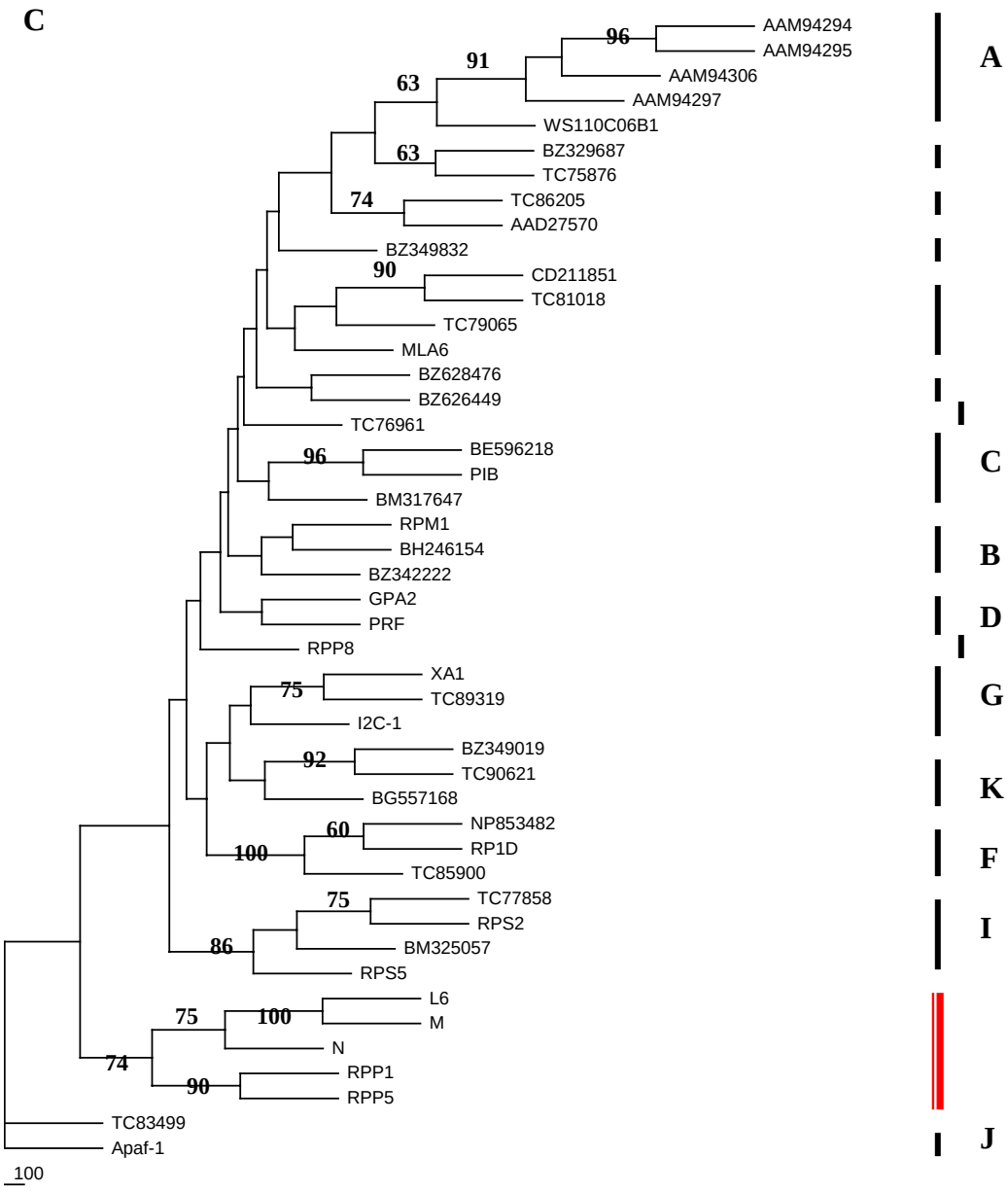


Figure 4. Continued.

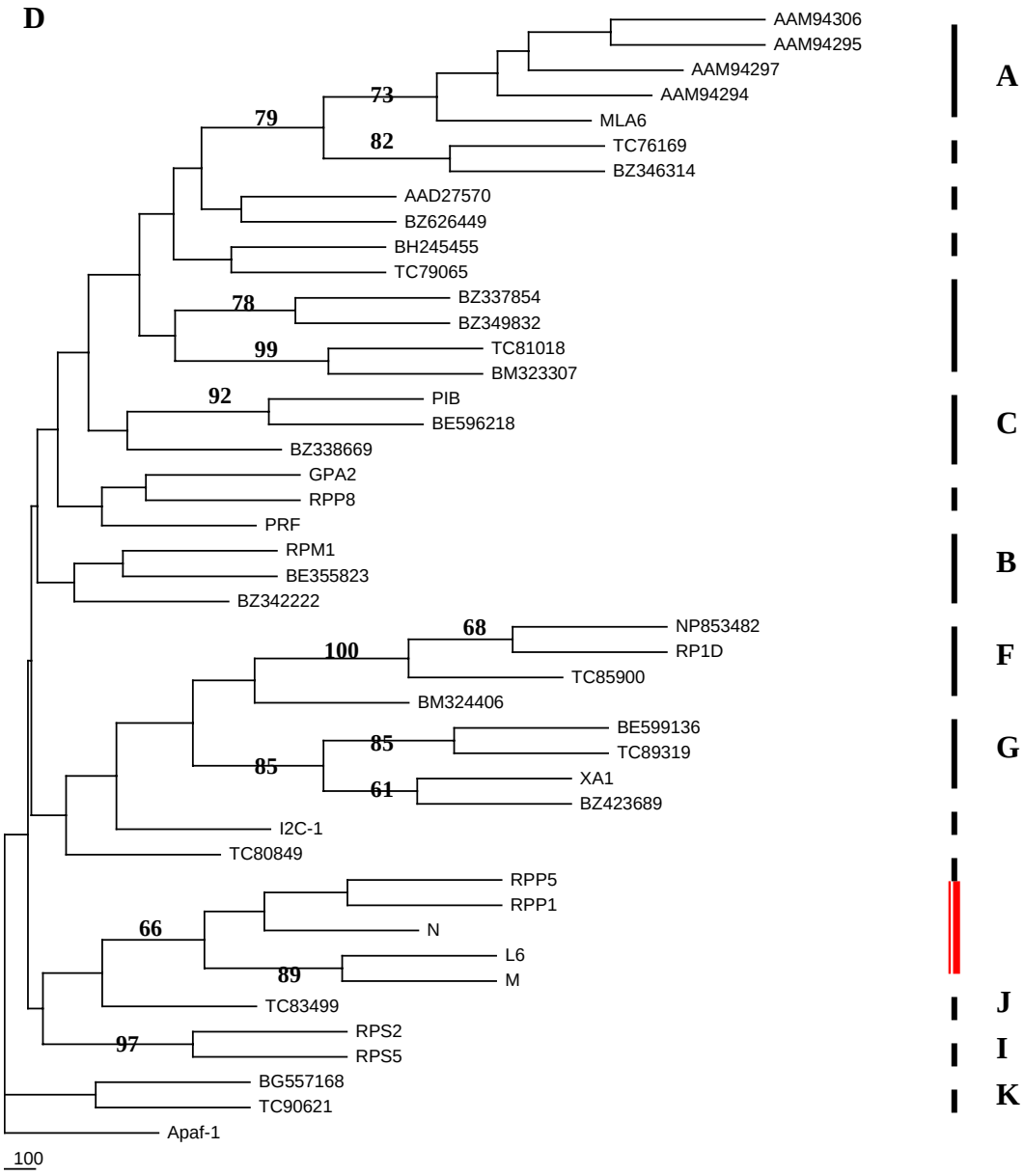


Figure 4. Continued.

(A to K) based on tree topology, some of which associated with the NBS domains of cloned R genes. The NBS family branches matched well with classified motif patterns. Two families (D, E) consisted of only one NBS sequences, while other families were each composed of several members. According to the previously defined distinction between the TIR class and the non-TIR class, all isolated sorghum NBS sequences seemed to belong to the non-TIR class type of R-genes.

Identification of Rice Orthologs of Sorghum NBS Sequences

To find orthologous sequences between *S. bicolor* and rice, we used a two-way sequence similarity comparison and phylogenetic methods to construct the evolution of these sequences more reliably. Previously, we classified sorghum NBS sequences into 11 phylogenetic groups based on phylogenetic topology related to 16 known R genes. Each group of sorghum NBS sequences was used to find the best hits of rice sequences at the nucleotide level. The best hits and the reciprocally best hits between sorghum NBS sequences and rice homologs are listed in Table 10.

The phylogenetic trees were analyzed to identify orthologous sequences between *S. bicolor* and rice expressed sequences. Two groups (A and F in Figure 4A) of sorghum NBS sequences and 10 rice sequences shown the best hits were aligned and used to calculate phylogenetic trees. Barley and maize homologs (five best hits each) were included into the phylogenetic tree to improve the chance of finding correct orthologs. After manual inspection of sequence alignments, rice, barley and maize homologs mismatched in the conserved motif region were removed for phylogenetic tree construction. The phylogenetic trees are calculated in different ways and the results are

Table 10. Rice Sequences Homologous to Sorghum NBS Sequences

NBS sequences[a]	Rice genes[b]	Map position	E-value[c]	Recip. best hits[d]
Sorghum NBS sequences in group A				
AAM94294	TC256094	Ch2: 11.4	7.8×10^{-113}	*
AAM94295	NP906632	Ch11: 19.2	1.3×10^{-115}	*
AAM94297	NP918495	Ch8: 4.0	2.8×10^{-107}	*
AAM94306	TC280286	-	6.9×10^{-99}	*
CD212839	NP895111	Ch12: 22.7	5.4×10^{-40}	
TC81885	TC266000	Ch11: 21.0	7.1×10^{-56}	
Sorghum NBS sequences in group B				
AW564339	TC279072	Ch11: 6.5	5.8×10^{-42}	
BE355823	TC271530	-	4.2×10^{-40}	*
BH246040	TC273025	-	2.0×10^{-33}	
BH246154	NP908519	Ch4: 26.9	1.5×10^{-42}	
BZ342222	TC278017	Ch1: 20.1	3.1×10^{-47}	
NP239121	TC260079	-	1.9×10^{-60}	*
Sb_RGA55	TC273026	Ch7: 4.6	5.3×10^{-31}	
Sorghum NBS sequences in group C				
BE596218	TC269539	Ch2: 34.6	9.7×10^{-52}	
BZ343608	TC255207	Ch6: 10.3	9.3×10^{-22}	
BZ367728	TC269539	Ch2: 34.6	1.7×10^{-29}	
NP239124	NP885584	Ch6: 2.3	1.0×10^{-24}	
TC76961	NP885584	Ch6: 2.3	2.5×10^{-95}	*
Sorghum NBS sequences in group D				
TC79065	TC279077	Ch1: 13.0	3.9×10^{-71}	*
Sorghum NBS sequences in group E				
BZ369917	TC255561	Ch3: 27.2	5.9×10^{-38}	
Sorghum NBS sequences in group F				
BM324406	TC266180	Ch5: 19.3	1.2×10^{-71}	
NP853482	TC280975	Ch1: 32.8	0.0	*
RHOH_13_F05.g1	TC257406	Ch1: 32.8	1.2×10^{-65}	
TC85900	TC257409	-	0.0	
Sorghum NBS sequences in group G				
BE599136	TC272697	Ch6: 28.8	1.5×10^{-34}	
BZ423689	NP932712	Ch4: 31.0	6.5×10^{-47}	
Sb_RGA75	NP931369	Ch4: 31.3	8.7×10^{-37}	
Sb_RGA80	TC255198	Ch8: 11.9	4.9×10^{-47}	
Sb_RGA130	NP898227	-	3.0×10^{-58}	
TC87218	TC266180	Ch5: 19.3	3.2×10^{-74}	*
TC89319	NP1100995	-	8.8×10^{-227}	*

Table 10. Continued.

NBS sequences	Rice genes	Map position	E-value[b]	Recip. best hits[c]
Sorghum NBS sequences in group I				
TC89312	TC272431	Ch4: 25.2	6.3×10^{-102}	*
Sorghum NBS sequences in group J				
NP239122	NP002099	-	3.4×10^{-49}	*
TC83499	TC251324	-	7.1×10^{-274}	*
Sorghum NBS sequences in group K				
BG557168	TC268224	Ch8: 26.9	1.6×10^{-54}	*
TC90621	TC281900	Ch3: 35.4	4.1×10^{-91}	*
Sorghum NBS sequences in unspecified group				
AAD27570	TC269622	Ch12: 22.7	5.7×10^{-161}	*
AW285775	TC282525	-	4.4×10^{-49}	
BM323307	TC279371	Ch10: 1.9	4.6×10^{-40}	
BH245455	TC262838	-	3.8×10^{-39}	
BH246056	TC270444	Ch11: 22.1	1.2×10^{-32}	
BZ337854	TC278712	Ch10: 1.8	1.2×10^{-58}	
BZ341506	TC270316	Ch10: 1.8	4.1×10^{-75}	
BZ346314	TC269622	Ch12: 22.7	3.0×10^{-32}	
BZ349832	NP895111	Ch12: 22.7	1.4×10^{-33}	
BZ626449	TC282525	-	8.8×10^{-38}	
BZ628476	NP655950	Ch7: 15.1	6.6×10^{-46}	
NP239123	TC281873	Ch3: 14.8	3.2×10^{-21}	*
TC75876	NP906632	Ch11: 19.2	8.4×10^{-50}	
TC76169	NP258957	Ch1: 23.5	6.2×10^{-55}	
TC79065	TC279077	Ch1: 13.0	3.9×10^{-71}	*
TC80849	TC266233	Ch2: 14.8	1.3×10^{-57}	
TC80927	TC265475	Ch8: 19.5	1.3×10^{-26}	
TC81018	TC279371	Ch10: 1.9	3.7×10^{-83}	*

[a]Sorghum NBS sequences are listed by group based on phylogeny of sorghum NBS sequences in Figure 4.

[b]Rice genes are the best hits found in BLAST searches from *Oryza sativa* Gene Indices (OsGI) at TIGR.

[c]E-value is based on nucleotide sequence level similarity.

[d]Asterisks (*) indicates that two homologous sequences are reciprocal best matches.

highly dependent on the chosen method and parameters used. In this study, three different tree-building programs (neighbor-joining, maximum parsimony, and maximum likelihood) and different models of evolution for neighbor-joining method to calculate different trees were analyzed for orthologs. Overall, five combinations of different trees were used to find orthology. Assignments were made only if a majority of programs supported the orthology with high confidence value. Table 11 lists the sequences involved in 8 putative sorghum-rice orthology assignments that were identified with the described procedure. The different types of orthologous relationships are illustrated in Figures 5 and 6.

BAC Screening of PCR Amplified NBS Sequences

For genome-wide scanning of NBS sequences in *Sorghum bicolor*, an entire BAC library of 13,440 clones (~ 3 X genomes) on ten high-density filters was screened by hybridization with eight NBS sequences isolated by PCR amplification. The similarity of RGA probe sequences was at maximum 59 % at the amino acid level. These represent three families of NBS sequences in sorghum (Table 4). The positive BAC clones can be detected with signals in two opposite dots. One probe (Sb_RGA75) hybridized to the single BAC clone (08G21). For all other probes except Sb_RGA75, multiple clones showing a positive signal were detected per each probe. Four probes (Sb_RGA80, Sb_RGA125, Sb_RGA181 and Sb_RGA182) hybridized to at least one shared sorghum BAC clone (21J23, 32M13 or 36I11) (Table 12). This suggests that the shared BAC clones may contain mostly the same insert fragment, or RGA probes have many copies

Table 11. Orthology Assignments between *Sorghum bicolor* and Rice NBS Sequences[a]

NBS Group[b]	Orthologs		
	S. bicolor orthologs	*O. sativa* phylogenetic orthologs	*O. sativa* blast best hits[c]
A (MLA)	AAM94294	NP906632	TC256094*
A	AAM94295	NP906632	NP906632*
A	AAM94297	NP906632	NP918495*
A	AAM94306	NP906632	TC280286*
A	CD212839	TC273139	NP895111
F (Rp1-D)	NP853482	TC277627	TC280975
F	RHOH_13_F05.g1	TC277627	TC257406
F	TC85900	TC277627	TC257409

[a]Two groups (A and F) of sorghum NBS sequences were further analyzed to find phylogenetic orthologs (see Materials and Methods).
[b]NBS group is based on phylogeny of sorghum NBS sequences in Figure 4.
[c]Asterisks (*) indicates reciprocal best hits between sorghum and rice NBS sequences.

A

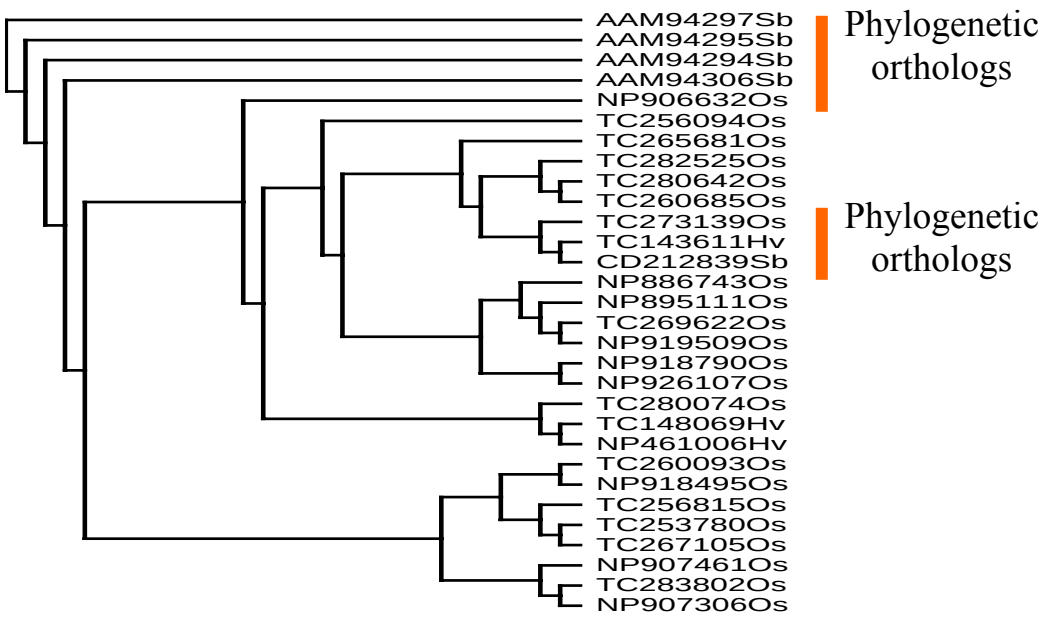
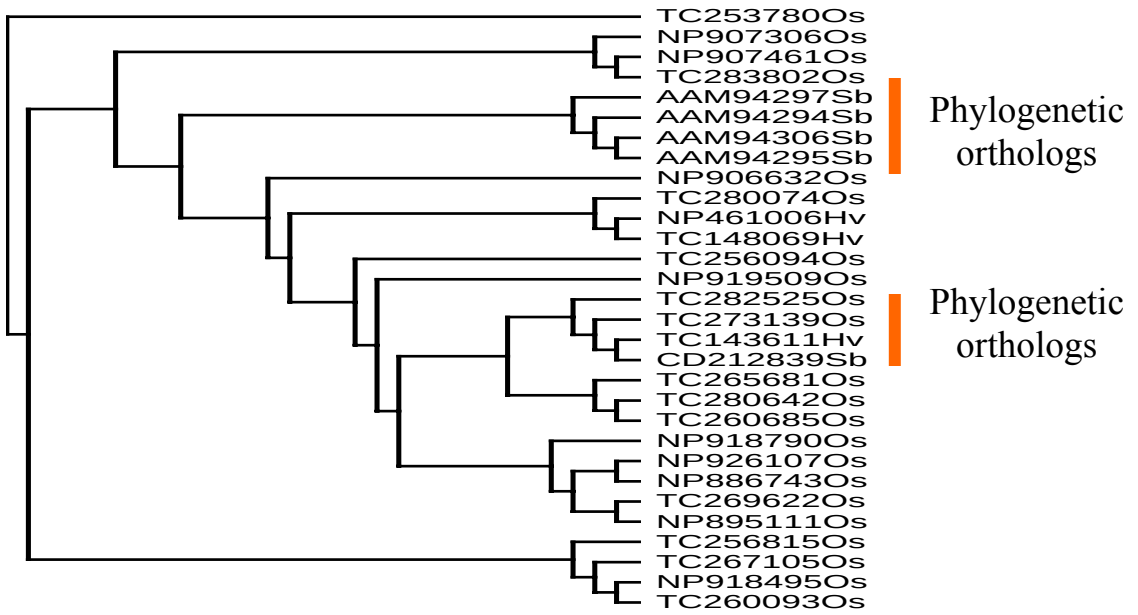


Figure 5. Rice Orthologs of Sorghum NBS Sequences (Group A). The trees are calculated using 5 different methods, with 100 bootstrap replicates: **A,** maximum likelihood (ML) tree with JTT (see Materials and Methods); **B,** maximum parsimony (MP) tree; **C,** neighbor-joining (NJ) tree with JTT; **D,** NJ tree with PAM; **E,** NJ tree with Kimura's distance. Sequences from different grass species are distinguished by the end two letters: Hv, barley; Os, rice; Sb, sorghum.

B



C

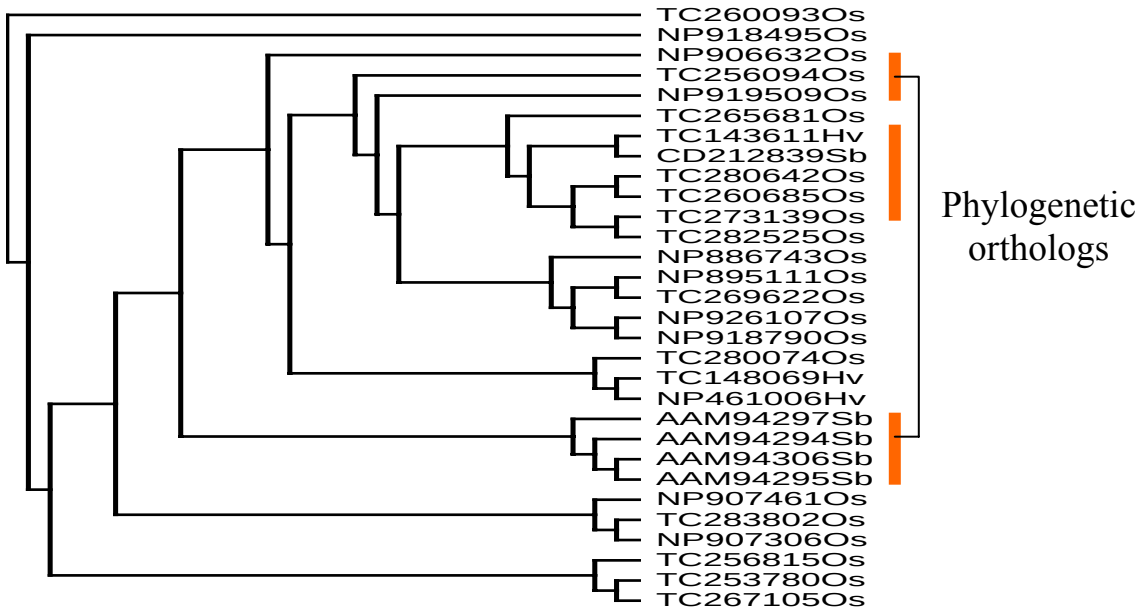
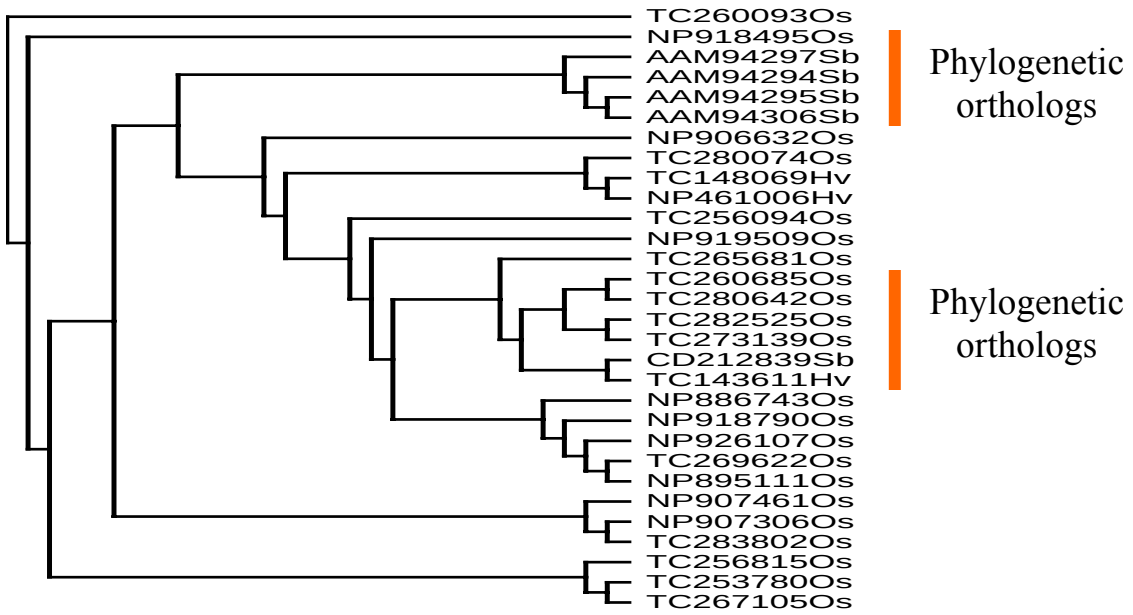


Figure 5. Continued.

D



E

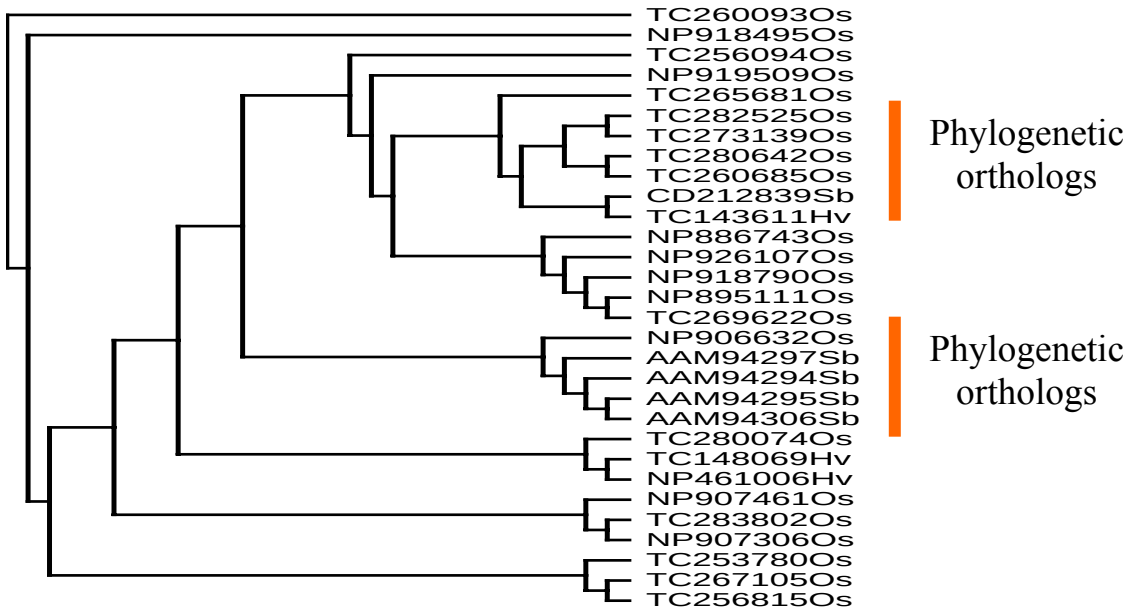


Figure 5. Continued.

A

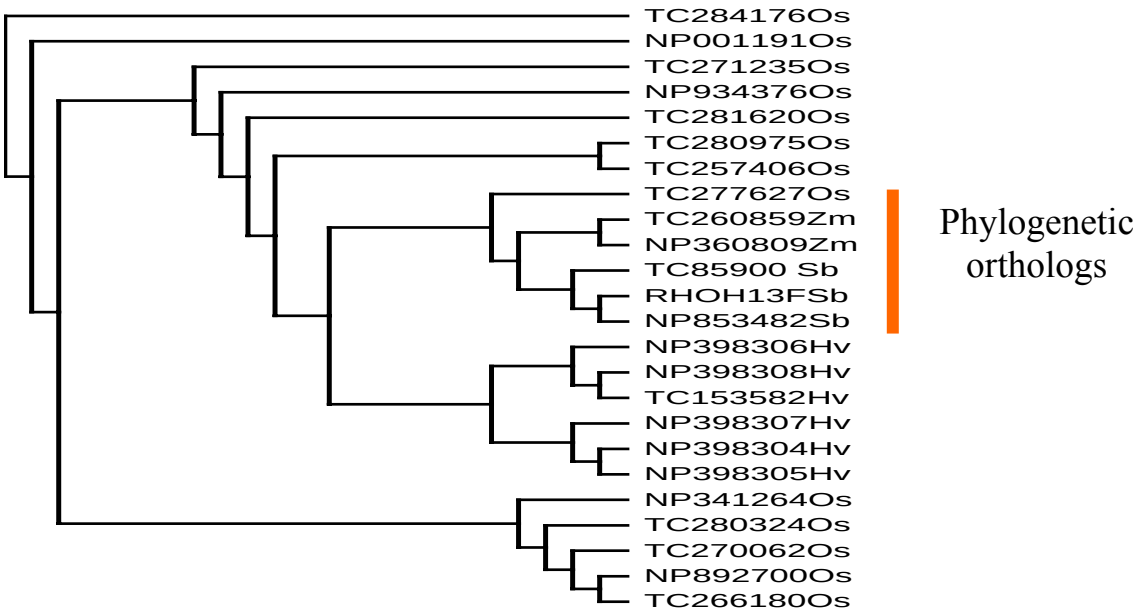
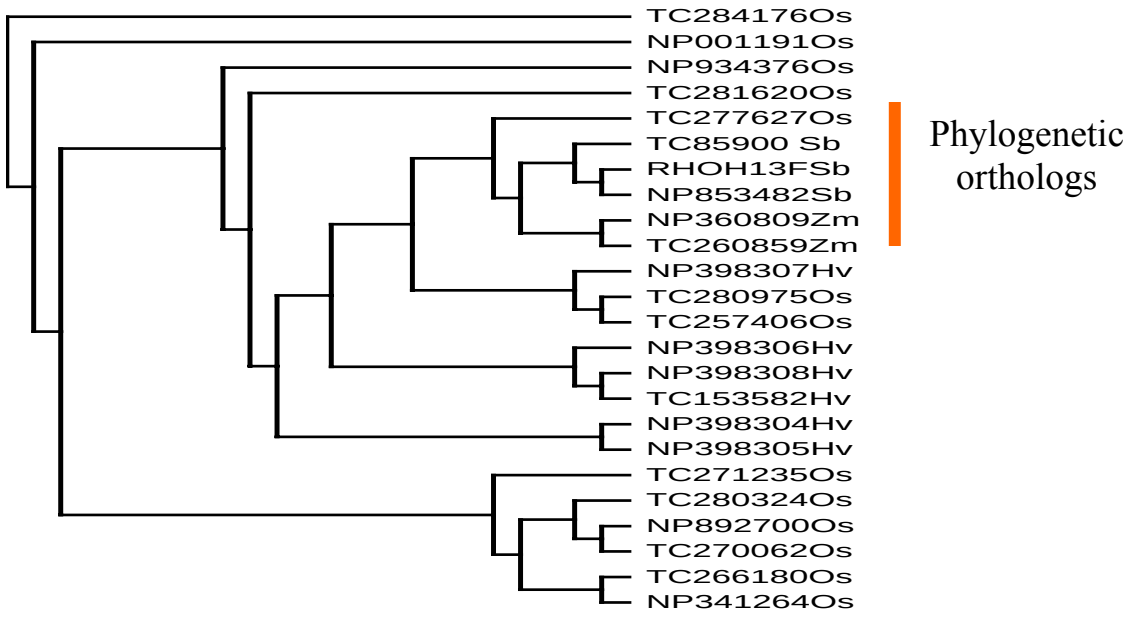


Figure 6. Rice Orthologs of Sorghum NBS Sequences (Group F). The trees are calculated using 5 different methods, with 100 bootstrap replicates: **A,** maximum likelihood (ML) tree with JTT (see Materials and Methods); **B,** maximum parsimony (MP) tree; **C,** neighbor-joining (NJ) tree with JTT; **D,** NJ tree with PAM; **E,** NJ tree with Kimura's distance. Sequences from different grass species are distinguished by the end two letters: Hv, barley; Os, rice; Sb, sorghum; Zm, maize.

B



C

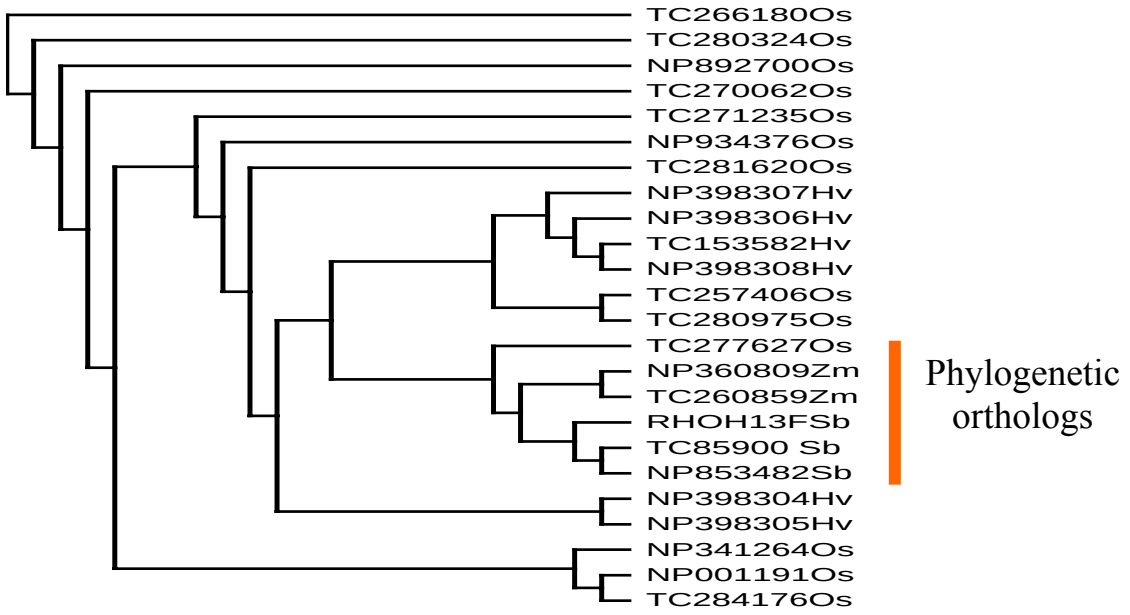
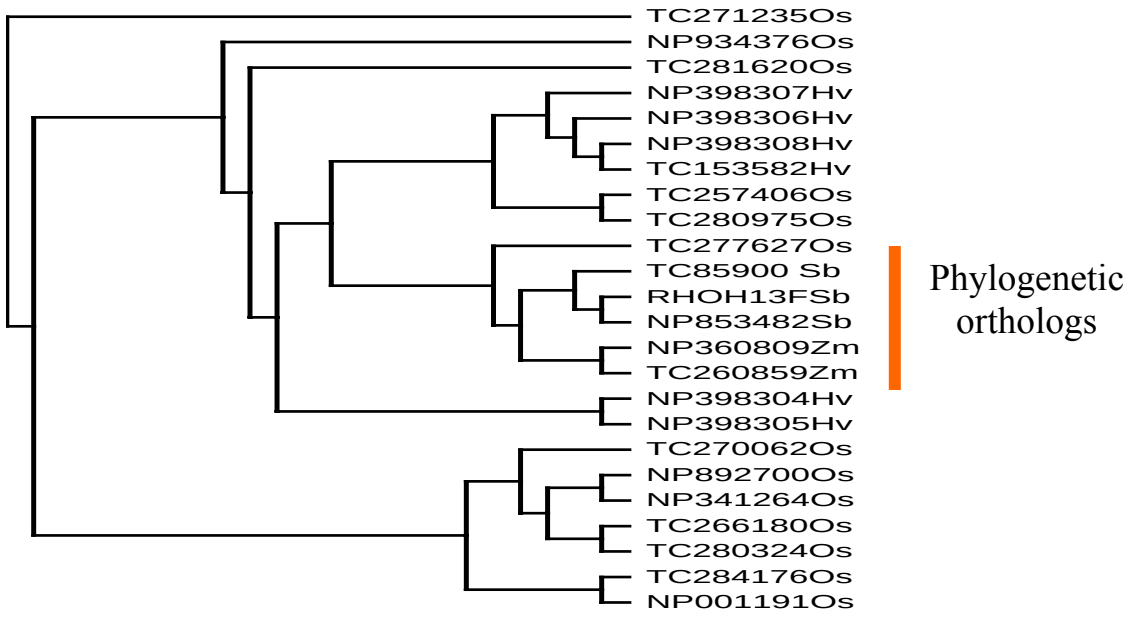


Figure 6. Continued.

D



E

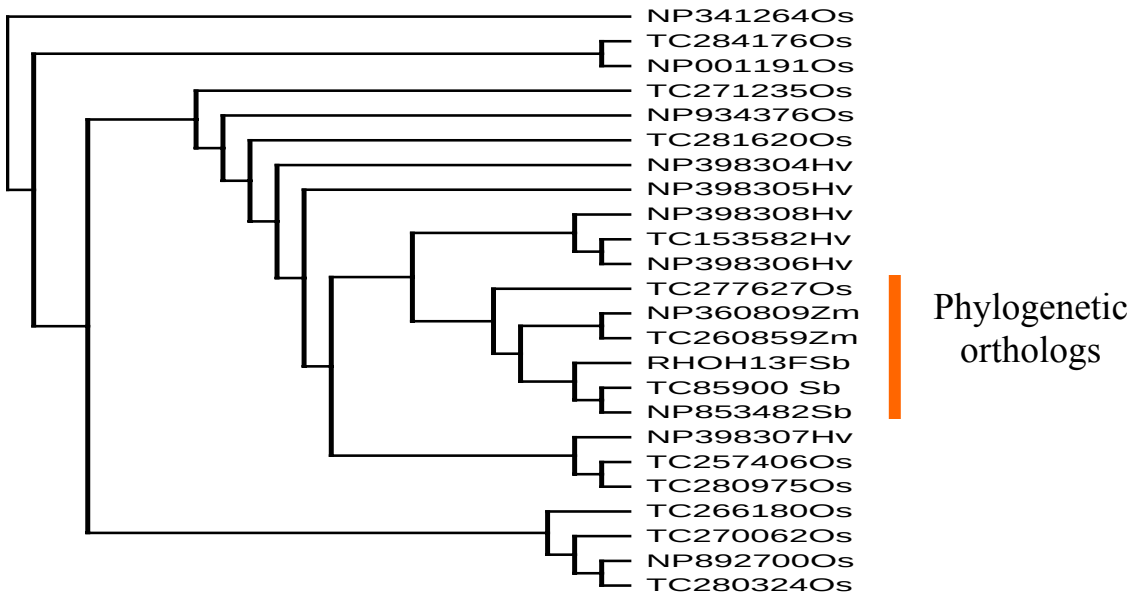


Figure 6. Continued.

Table 12. Sorghum BAC Clones Hybridized with PCR Amplified RGA Sequences

RGA probes[a]	BAC clones hybridized to RGA probes[b]																	
Sb_RGA50																	37 B 11	37 I 11
Sb_RGA55			16 H 20	17 L 01	17 M 11			30 E 16	32 H 06									
Sb_RGA75	08 G 21																	
Sb_RGA80							21 J 23					35 N 07		36 M 06				
Sb_RGA125							21 J 23			32 M 13			36 I 11					
Sb_RGA130						18 M 15					33 I 09						37 B 11	37 B 12
Sb_RGA181		13 D 23								32 M 13			36 I 11					
Sb_RGA182		13 D 23								32 M 13			36 I 11					

[a]RGA clones were digested with *EcoRI* to isolate inserts from vector sequences, and then radiolabelled by random priming method (*Ready-to-go*® DNA radiolabelling kit).
[b]BAC clones shown signals in two opposite dots were considered as positive clones.

in the genome like multigene family members.

Detection of the Restriction Fragment Length Polymorphism

After initial screening with seven restriction enzymes (*Bam*HI, *Eco*RI, *Eco*RV, *Hind*III, *Pst*I, *Xba*I and *Xho*I) four restriction enzymes - *Eco*RI, *Eco*RV, *Hind*III and *Xba*I - were used to test for RFLPs in the mapping parents because these enzymes detected polymorphism more frequently than the other three enzymes. Of eighty-nine NBS sequences from several sources (EST clones, clones of genomic DNA fragments or from PCR generated products), fifty-five sequences were identified for use in mapping. Each of these could be easily amplified by PCR or represented clones for which the insert fragment could be readily isolated after digestion, and separation from the vector sequence. The similarity among all these probe sequences was below 75%. Thirty-two (58.2 %) revealed a RFLP between the two parental lines with at least one of the four restriction enzymes. The RFLP frequency detected with each restriction enzyme is shown in Table 13. Typical RFLP band patterns are shown in Figure 7. Twenty-nine probes (52.7 %) detected only a single fragment per parent; whereas twenty-six probes (47.3 %) detected 2 or more fragments per parent with each of the four restriction enzymes. Overall, the average number of fragments detected/per probe was 1.43, and the range of the average among the four restriction enzymes was from 1.3 to 1.6 (Table 13).

Mapping of the NBS Sequences

The ten NBS probes that showed distinct and easily detectable polymorphic band were

Table 13. Polymorphism Levels between BTx623 and IS3620C Detected by Sorghum NBS Sequences Using Four Restriction Enzymes[a]

Restriction Enzymes	% polymorphism released	Cumulative RFLP (%)	Number of fragments detected/probe
*Eco*RI	32.7	32.7	1.6
*Eco*RV	27.3	41.8	1.3
*Hind*III	25.5	50.9	1.3
*Xba*I	32.7	58.2	1.5

[a]Based on the analysis of 55 sorghum NBS sequences

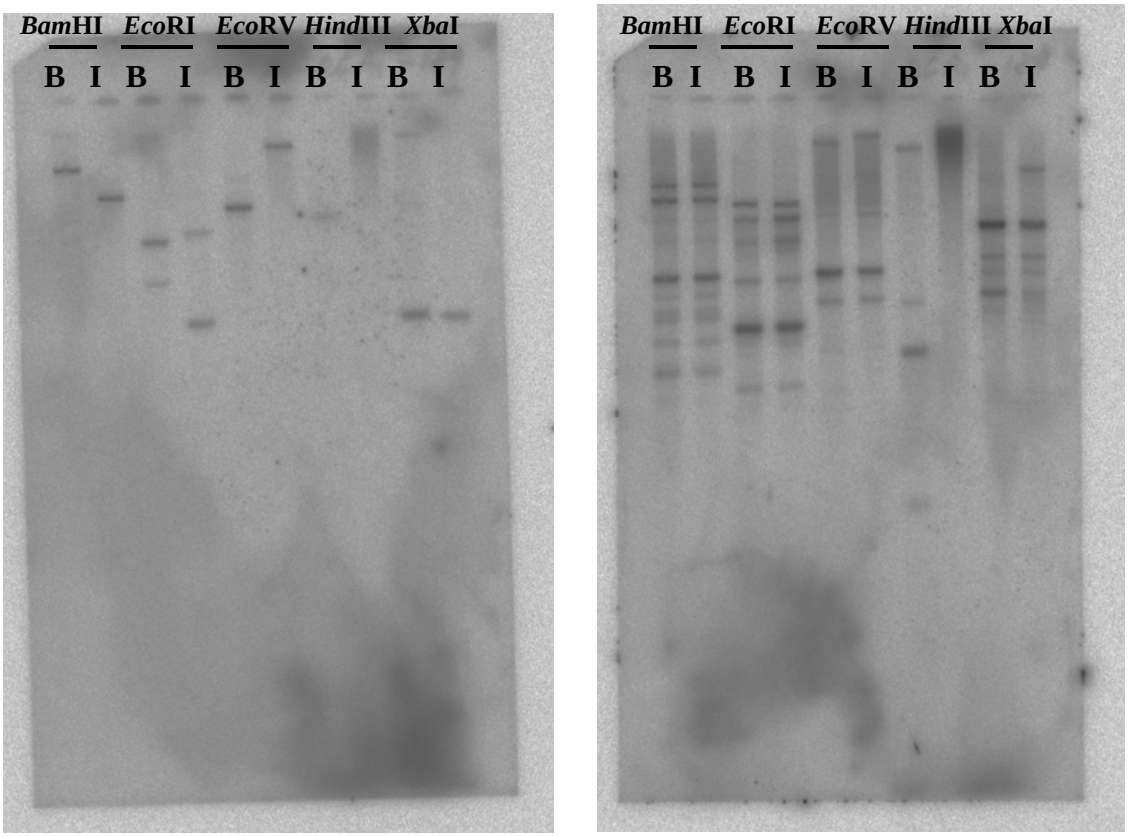


Figure 7. Restriction Fragment Length Polymorphism (RFLP) Analysis of Genomic DNA from Sorghum Parental Lines (B, BTx623; I, IS3620C). Hybridization using sorghum NBS sequence probes [BH245455 (left) and BZ423246 (right)] revealed single or multiple band patterns with/without polymorphisms.

used for further analysis. Among these NBS probes, eight hybridized to a single fragment from one parent and 1 or 2 from the other parent and two hybridized to 2 fragments from both parents.

Genetic mapping placed NBS sequences on one linkage group and six single loci. One NBS cluster representing three of the different classes spanned a distance of 39.536 cM containing four NBS probes (Figure 8). The position of one (BH245455) of these probes was previously reported to be on linkage group H on another high-density genetic map, which was constructed using a highly polymorphic mapping population from the cross *Sorghum bicolor* X *S. propinquum* (Bowers et al., 2003). Three more NBS sequences - BH246056, AAM94294 (shown as AF527807 in Figure 9) and AAM94319 (shown as AF527809) – were linked to the BH245455 locus within 20 cM in this linkage group H. However, we didn't find any polymorphism among these NBS sequences for further comparison. The NBS loci that distributed into linkage group H are indicated with arrows in Figure 9.

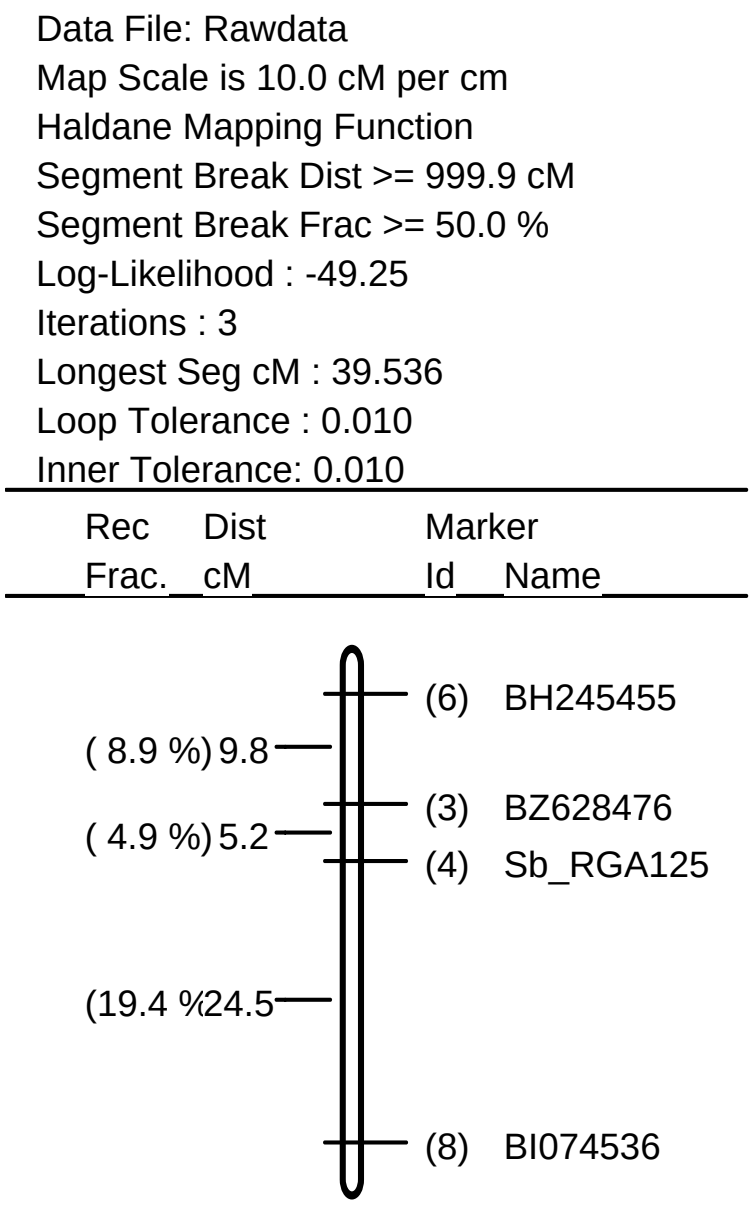


Figure 8. A Linkage Group Mapped with Four Sorghum NBS Sequences.

Figure 9. Distribution of NBS Sequences on the Linkage Group (LG) H of a High-Density Genetic Map Constructed Using the Population from Interspecific Cross *Sorghum bicolor* and *S. propinquum* (Bowers et al., 2003).

DISCUSSION

Sorghum NBS Sequences Are Non-TIR Specific

It has been suggested from analysis of RGA sequences from rice, barley, maize and wheat that monocots lack a family of NBS-LRR genes with a TIR motif (Pan et al., 2000; Bai et al., 2002; Quint et al., 2002; Madsen et al., 2003). Although TIR domains are found in their genomes, they are not associated with NBS or LRR regions (Zhou et al., 2004). But in Arabidopsis, a model dicot plant for which the entire genome sequence is available, two thirds of the NBS-LRR genes identified contain TIR domains (Meyers et al., 2003). Sorghum, one of the major cereal crops and a member of the grass family (Poaceae), has also been included as a non-TIR monocot. However, there were only five PCR products used for the sorghum samples that Meyers et al. (1999) used to reach this conclusion. In this study, we tried to verify their conclusion by collecting and analyzing a large number of sorghum NBS sequences.

As expected from many previous reports, only non-TIR specific NBS sequences could be detected in sorghum and no evidence for TIR-NBS sequences was found. The initial test was done using PCR amplification. Using two subgroup-specific degenerate primers matching two motifs that are conserved among the NBS regions of R genes or R gene homologs, we were able to contrast the ability to PCR amplify products using non-TIR specific primers with the inability to do so using TIR-specific primer combinations. Consistently, no primer combination with at least one TIR-specific backward primer could be used to amplify PCR products from sorghum genomic DNA. This suggests that

while the primer sequences used in this study can not represent all TIR-specific sequences, there might be no TIR-specific motifs in the sorghum genome. Pan et al. (2000) used the same primer sets to amplify PCR products from tomato and wheat genomes, and failed to obtain the products from wheat using TIR-specific primers. Bai et al. (2002) who performed extensive PCR amplifications with many primers was able to isolate NBS sequences from rice, but also noted the lack of TIR-specific sequences in the genome. This observation can also be supported by several other studies that used PCR strategy to obtain NBS sequences from monocot plants (Meyers et al., 1999; Bai et al., 2002; Madsen et al., 2003; Irigoyen et al., 2004).

Searching for TIR-specific NBS sequences from public molecular databases has become a useful method of identifying this class of RGAs as a result of the tremendous amount of sequence data (whole genome sequences in some model plants) that has become available in molecular databases. Furthermore, the proper searching tools for extracting homologous sequences have been developed that enhance the use of this strategy (Eddy, 1998). In the case of sorghum, although whole genome sequencing projects are not yet underway, large-scale EST libraries have been launched in University of Georgia. Sequence homology searches (BLAST) identified 84 NBS sequences from the available sorghum molecular databases which have grown rapidly during last 3 years. When searched with TIR-specific NBS domains of known TIR-NBS-LRR R genes, no homologous sequences were detected, or if detected, they (5 sequences detected when used RPP5 NBS domain as a query) were at the lower ″expect value″, right above the cutoff limit (0.0001 used here). But all these sequences were non-TIR

specific and were detected with the higher expect value when searched using non-TIR-NBS-LRR R genes. Recent genome-wide analyses of NBS-LRR genes in Arabidopsis and rice adopted the missing-error-minimized searching strategy to find even distantly related sequences, demonstrating the absence of TIR-NBS-LRR genes in the rice genome compared to the dominant number (two thirds) of TIR-NBS-LRR genes in Arabidopsis (Meyers et al., 2003; Zhou et al., 2004). PCR products and NBS sequences collected from databases were further analyzed. Motif structures in the NBS domain further confirmed their non-TIR group specificity. A detailed list of motif structures limited to subgroup-specific sequences has been established from previous articles (Meyers et al., 1999; Pan et al., 2000). Eight major motifs have been identified in the NBS domain and some of them could be effectively used to distinguish TIR-specific sequences from non-TIR-specific sequences. These diagnostic motifs were called Kin-2, RNBS-A (RNBS-I in rice) and RNBS-D (RNBS-V in rice) (Meyers et al., 1999). Our data showed the existence of the eight major motifs with slightly different consensus sequences. All are typical of non-TIR-specific motif appearances when compared consensus sequences of diagnostic motifs. A few variations were observed, but these did not resemble characteristics of typical TIR-specific motifs. TIR domains are thought to function in signal transduction (Ellis and Jones, 1998), and may also be involved in pathogen recognition (Ellis et al., 1999). The absence of TIR domains in sorghum suggests the loss of TIR-related defense signal pathways.

The phylogeny of sorghum NBS sequences further supported the previous conclusion and the above observation that no TIR-specific sequences exist in the

sorghum genome. The phylogenies of TIR- and non-TIR-specific sequences were clearly distinguished: members of the grass family were only found in non-TIR types of branches (Cannon et al., 2002). Sorghum NBS sequences were all branched to the non-TIR types of branches and a TIR type branch composed of TIR-specific R genes (N, M, L6, RPP1 and RPP5) was distinguished and contained no sorghum NBS sequences.

Sorghum NBS Sequences Are Diverse and Abundant

The NBS-encoding sequences isolated from sorghum showed considerable sequence variation. The nucleotide sequences of the collected NBS domains were aligned to each other using the program 'Sequencher'. The number of NBS sequences was reduced by only two (only two contigs further detected) when the similarity cutoff value was decreased from 95% to 65%. Moreover, phylogeny of sorghum NBS sequences showed closely clustered nodes and long-branch lengths, suggesting high divergence of these sequences. Based on topology of the tree containing known R genes, sorghum NBS sequences were classified into 11 groups where each group (except groups - H, J and K) includes at least one known R gene sequence. The similarity range among inter-group members is low and the maximum value of similarity did not reach 60% even within group members. In fact, the various sorghum NBS sequences that were identified showed strong sequence similarity with almost all known non-TIR-type R-genes. These results provide further evidence that TIR-type sequences are absent not only in the whole of rice genome (Meyers et al., 1999; Pan et al., 2000; Zhou et al., 2004) but in other cereal genomes as well.

NBS sequences are abundant in plant genomes. For instance, the Arabidopsis genome is estimated to contain approximately 200 NBS-encoding genes (150 of the TIR type and 50 of the non-TIR type) (Meyers et al., 2002, 2003). The rice genome contains 535 NBS-coding sequences, including 480 non-TIR NBS-LRR genes and no TIR-NBS-LRR genes (Zhou et al., 2004). Obviously, sorghum may have significantly greater numbers of R-genes than were revealed here. The Genome Sequence Survey (GSS) database was used to predict the number of NBS-LRR genes in sorghum. As of March 2003, GSS database consists of 35,910 genomic sequences that were approximated to include 2.04×10^7 bases in gross size. Because only 569 bp from each sorghum entry from a methyl-filtered shotgun library are estimated to be high quality sequence, we used this value for the number of base pairs analyzed. When using an estimated genome size of 7.5×10^8 bp for sorghum, the available high-quality reads would represent 2.7 % of the sorghum genome. Since genomic sequences of the GSS database are created from methyl-filtered shotgun genome library (www.ncbi.nlm.nih.gov). We re-estimated an effective genome size of 3.75×10^8 bp for sorghum, assuming that half of the sorghum genome ($\approx 3.75 \times 10^8$ bp) contains non-coding repetitive sequences. The 50 % figure is based on Cot-based sequence analysis of the sorghum genome (Peterson et al., 2002). Thirty-five NBS sequences were identified from the entries of methyl-filtered shotgun library in GSS database (Table 6), and all were of non-TIR type. This suggests that there are about 644 non-TIR NBS-encoding sequences in the sorghum genome. This would represent 1.7 % of all sorghum genes if it is assumed that all unique sequences (37,232 entries) in the *Sorghum bicolor* Gene Index (SbGI) represent different sorghum genes.

Finding Rice Orthologs

Orthologs are likely to have the same functions and similar biological roles. Thus, these orthologs might be an invaluable source for clarifying the function of uncharacterized genes. If two orthologs had a common ancestor and directly diverged by speciation only, they should be easy to recognize as the most similar sequences in a two-way (reciprocal) sequence similarity comparison. However, complicating factors such as subsequent gene duplication and different divergence rates make the simple two-way sequence comparison technique unreliable. Therefore, we used phylogenetic methods as well as two-way blast to find orthologous sequences between sorghum and rice.

Sorghum NBS sequences were classified into 11 groups in which at least one known R gene from another species was included. Each group of sequences was then used to query the most homologous sequences in *Oryza sativa* Gene Indices (OsGI) at TIGR. The groups of sorghum NBS sequences and their rice homologs (the best hits) are described in Table 10. The most abundant category of rice homologs is NBS-LRR-like protein or putative disease resistance proteins. Known R genes (*Pib* and *Xa1*) were found as the best hits of sorghum NBS sequences (BE596218, BZ367728 and TC89319). All sorghum NBS sequences matched rice homologous sequences at e-values lower than 3.2×10^{-21}. When searched with rice homologs against *Sorghum bicolor* Gene Indices (SbGI), twenty sorghum NBS sequences were found as reciprocally ″best hits″.
The phylogenetic trees were analyzed to identify orthologous sequences between sorghum NBS sequences and rice homologs. All sorghum NBS sequences from a given group (two groups analyzed in this study) and ten rice homologs were aligned and used

to calculate phylogenetic trees. Other barley and maize homologs were included into the phylogenetic tree to improve the chance of finding true orthologs. Because the phylogenetic trees can be calculated in different ways and the results are highly dependent on the chosen method, we used five different methods (see Materials and Methods) to calculate different trees (Figure 5 and 6). Table 11 lists the sequences involved in orthology assignments that were identified with the phylogenetic analysis.

Sorghum NBS-LRR Genes Were Clustered and Non-randomly Distributed in the Genome

Genome-wide molecular data clearly demonstrated that plants have R-genes arrayed in complex clusters (Meyers et al., 2003; Zhou et al., 2004). Indeed, clustering of R-genes and homologous sequences may facilitate the generation of diversity and new resistance specificities. Similarly, NBS-LRR genes are distributed unequally in the plant genome. As was observed in Arabidopsis and rice, one or two chromosomes contain dense distribution of NBS-LRR genes which are found in characteristic clusters. Some of these clusters consist of single genes or a diverse family of NBS-LRR gene sequences (Meyers et al., 2003; Zhou et al., 2004). Sorghum NBS sequences showed a clustered distribution on the linkage group. Although a small number of sorghum NBS probes were used for mapping analysis, one cluster contained one third of the probes randomly collected in this study, suggesting that this linkage group may be one of the densely-distributed NBS containing chromosomes in sorghum. The clustering of sorghum NBS sequences is also demonstrated by BAC screening analysis. In this study, we screened the BAC clones by

probing with PCR amplified RGA sequences and found that three different classes of RGA probes hybridized to the same sorghum BAC clones. This suggests that these RGA sequences are located in a cluster within the contiguous region of this BAC. Because one of three RGA probes (Sb_RGA125) that hybridized to this BAC contig is also mapped into the cluster identified in this study, the shared BAC clones may be located within the mapped cluster.

The cluster revealed in this study is related to linkage group (LG) H (as defined by Peng et al, 1999; or the same as LG E as defined by Tao et al., 1998). One RGA probe (Sb_RGA125) in this cluster maps less than 20 cM from BH245455 which is mapped into linkage group H in a sorghum high-density genetic map (Bowers et al., 2003). Two sequenced RFLP markers (AF527807 and AF527809) that were found to contain RGA homology in this study also mapped to the same region (20 cM from BH245455 on Bower's map) (Bowers et al., 2003), but did not uncover polymorphisms between the mapping parents used here. Their predicted amino acid sequences show homology to *Mla1*, a gene which confers resistance to powdery mildew fungus in barley (Zhou et al., 2000; Song et al., 2002). Moreover, homologs of the maize rust resistance gene *Rp1-D* and major rust resistance QTL (quantitative trait loci) are also associated with this LG H in sorghum. Ramakrishna et al. (2002) identified ten bacterial artificial chromosome (BAC) clones from the sorghum BTx623 BAC library that hybridized to a probe *Rp1-D* gene (Collins et al., 1999). The BAC clones were physically mapped into a 350-kb contiguous region and contained five *Rp1* homologs in a 27-kb region in this contig map. Most of the sorghum BACs harboring *Rp1* homologs mapped close to

marker bnl3.04 on sorghum linkage group H. Markers umc130, and rz561 as well as bnl3.04 which are near *Rp1* in maize have been mapped and shown to flank the *Rph* (*Rp1-D* homologous) region of sorghum linkage group H (Wilson et al., 1999; Klein et al., 2000; Ramakrishna et al., 2002). Recently, McIntyre et al. (2004) mapped a rust resistance QTL onto linkage group H (they referred to LG E) of sorghum. This suggests that linkage group H may be a candidate of the region for a highly concentrated distribution of NBS-LRR genes in sorghum.

The linkage group H is more or less related to chromosome 11 in rice, on which *Rp1* homologous sequences are mapped (Ramalingam et al., 2003). The rice orthologs are located in similar order to matching sorghum NBS sequences (Figure 10). Ramalingam et al. (2003) mapped candidate defense genes in rice by using resistance gene homologous sequences as probes, and *Rp1* homologous sequences mapped to chromosome 11 on this map. Moreover, we tried to find rice orthologous sequences of sorghum NBS sequences, because orthologs are likely to have the same biological function and finding rice orthologs is an excellent starting point for comparative studies in sorghum. The recent drafts of the complete genome sequences in rice (Goff et al., 2002; Yu et al., 2002), which is a model plant and a member of grass family, allowed establishment of the chromosomal location of the NBS-LRR genes in the genome. The prediction of orthologs is quickly done by use of pair-wise-similarity detection programs. However, evolutionary events such as gene duplication and different divergence rates often make this similarity-based comparison unreliable. We used a more careful approach using phylogenetic methods, which is more consistent with the definition of

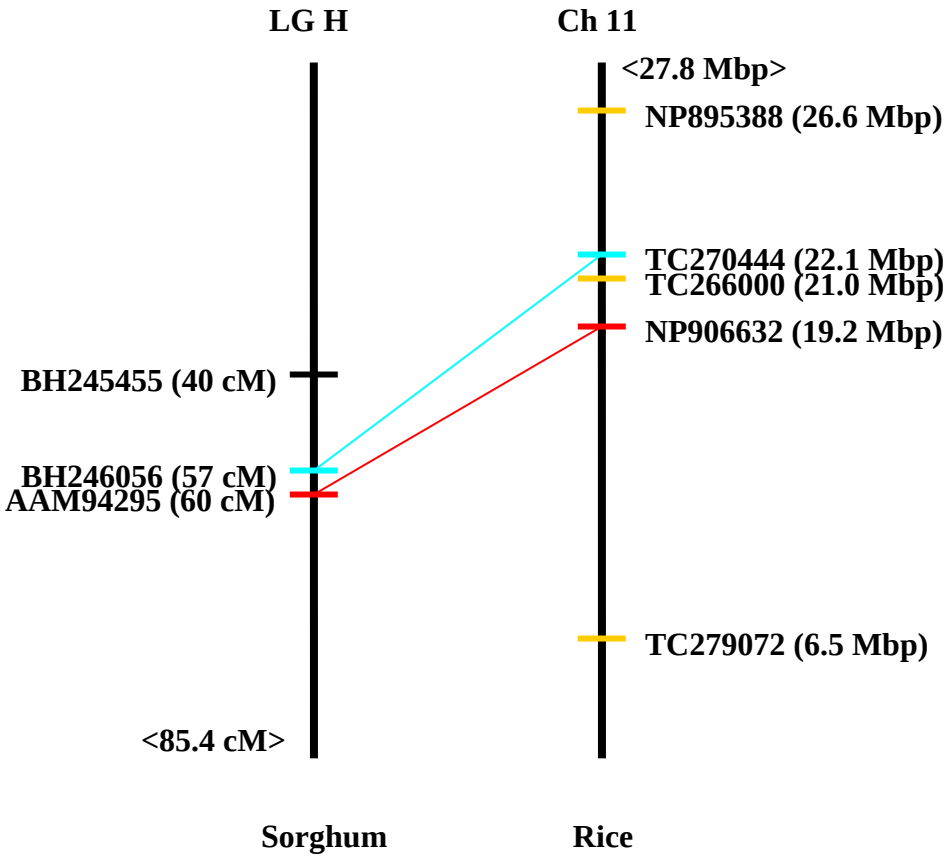


Figure 10. Comparison of Map Location between Sorghum NBS Sequences and Rice Homologous Sequences. Sorghum linkage group (LG) H is based on Bower's genetic map (2003). Rice chromosome 11 is reversely drawn by starting at the bottom. Map position bars of the same color indicate orthologs found in this study. The map location of sorghum counterparts of other rice homologs (yellow bar) is not yet determined.

orthology (Fitch, 1970). The results of this study show that the orthologs predicted by BLAST are often different from tree-based orthologs. Tree-based orthologs are more feasible than BLAST-predicted orthologs when the map location is included (Figure 10). Tree-based orthologs show syntenical relationship, but BLAST-predicted orthologs are variable in chromosomal location (Table 10). However, this conclusion must be considered tentative until the chromosomal locations of NBS-LRR genes are completely elucidated in sorghum and comparatively analyzed between sorghum and rice.

CHAPTER III

SUMMARY

Disease resistance (R) genes that confer resistance to a wide range of plant pathogens have been cloned and characterized from many plant species. Most cloned R genes (except for *Hm1* from maize and *Mlo* from barley) seem to code for components of signal transduction pathways. In addition to several R genes (e.g., *Pto*, *Xa21*, and the *cf* family of R genes) that encode receptor-like kinase and/or leucine rich repeat (LRR) domains, the majority of cloned R genes encode proteins with an N-terminal nucleotide-binding site (NBS) and a C-terminal leucine-rich repeat (LRR) region. Genes encoding NBS-LRR containing proteins are one of the most prevalent classes in plant genomes, comprising an estimated 1 % of all genes in Arabidopsis and in rice. Sequence motifs indicate that they act at the beginning of signaling pathways. Even though little is known about their function except disease resistance, they may also be involved in other aspects of plant biology including development and response to the environment.

The NBS-LRR class of R genes can be further subdivided into two groups based on the motif structure of the N-terminus of the predicted protein. The first group, termed TIR NBS-LRR, encodes an N-terminus with homology to the intracellular domains of the *Drosophila* Toll and the mammalian interleukin-1 receptor (TIR). The second group, termed non-TIR NBS-LRR, does not encode a TIR domain, but most members of this group instead encode a putative coiled-coil (CC) domain in their N-terminus. TIR and non-TIR NBS-LRR R genes can also be distinguished by the amino-acid motifs found within the NBS domain itself. Detailed comparisons of aligned NBS sequences reveal

several group-specific consensus sequences that can clearly distinguish two subfamilies. These motifs are so diagnostic that group-specific primers could be designed from these motifs and allow selective amplification of NBS sequences from either one of the two groups. Furthermore, the TIR and non-TIR NBS-LRR R genes also appear to be distinguishable functionally by involvement in different signal transduction pathways, which suggests a role of the N-terminal TIR or CC domains and/or related NBS motifs in the bifurcation of signaling pathways leading to plant resistance. In addition, database searching and experimental procedures revealed that the non-TIR group seems to be widely distributed in both monocot and dicot species, whereas the TIR group appears to be found exclusively in dicot species. The distinct distribution of these R genes among monocots and dicots indicates an ancient divergence of these two groups of genes in the plant genome.

Sorghum bicolor is an important species that is often used for studying comparative grass genomics and a potential source of beneficial genes for agriculture. However, there has been little interest in using *S. bicolor* as a target genome to clone the NBS-LRR genes. The study done here includes cloning, sequencing, database searching, and genetic mapping of sorghum non-TIR related NBS sequences within the NBS-LRR gene family in *S. bicolor*. We examined the map position of NBS sequences and the sequence diversity in *S. bicolor*. These studies may help to isolate new R genes and search for selectable markers for disease resistance, as well as answer to questions about evolution among resistance genes.

Resistance gene analogs (RGAs), especially NBS-encoding sequences, were

primarily identified by database searches, with others added from PCR products. We tested two subgroup-specific degenerate primers to know whether TIR-specific primers could amplify PCR products from sorghum genome or not. As expected from previous reports, no PCR products were amplified from TIR-specific primers which suggested the absence of TIR-NBS-LRR sequences in the sorghum genome. This observation was further supported by the result of database searching. In total, 84 sorghum NBS sequences were found from sorghum molecular sequences deposited to public databases, and all those sequences showed non-TIR specificity from the analysis which was done by both structural and phylogenetic methods. Thus, 89 sorghum non-TIR-specific NBS sequences including PCR amplified products were identified, and this number is estimated to be about 1/10 equivalents of all NBS-encoding sequences in the sorghum genome.

Sorghum NBS sequences contained eight major conserved motifs in the NBS domain and some of them showed only non-TIR specific type of consensus sequences. In addition to these major motifs, two additional motifs (RNBS-IV and RNBS-VI: found in rice, but not in Arabidopsis) were found between GLPL and MHDV motifs. Several minor motifs in the NBS domain found by MEME were variable enough to classify sorghum NBS sequences into 11 groups, and each group showed different motif pattern from other groups of sequences. The NBS sequences in each motif pattern group mostly belonged to single phylogentic groups in which at least one known R gene was included (two phylogenetic groups contain no known R genes). The phylogeny of sorghum NBS sequences showed the characteristic topology of NBS-LRR genes: clustered nodes and

long-branch lengths. The branch containing TIR-specific R genes was distinguished from branches of sorghum NBS sequences, suggesting sorghum NBS sequences are diverged from TIR-NBS-LRR genes.

Sorghum NBS sequences seem to be unevenly distributed through the genome. Four of ten probes randomly selected were mapped to one linkage group, while the others were mapped singly. Moreover, *Mla*, *Rp1-D* homologous sequences and quantitative trait loci (QTL) that contribute to rust resistance map to this linkage group. This linkage group is also related to chromosome 11 of rice which also has a high concentration of NBS sequences. Rice orthologous sequences of sorghum NBS sequences, which were mapped to the linkage group found in this study, and *Rp1* homologs are also placed on chromosome 11 in rice. NBS-LRR genes in sorghum are likely to be concentrated on the equivalent chromosome of sorghum.

REFERENCES

Aarts, N., Metz, M., Holub, E., Staskawicz, B.J., Daniels, M.J., and Parker, J.E. (1998). Different requirements for *EDS1* and *NDR1* by disease resistance genes define at least two *R* gene-mediated signaling pathways in *Arabidopsis*. Proc. Natl. Acad. Sci. USA **95,** 10306-10311.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein databases search programs. Nucleic Acids Res. **25,** 3389-3402.

Anderson, P.A., Lawrence, G.J., Morrish, B.C., Ayliffe, M.A., Finnegan, E.J., and Ellis, J.G. (1997). Inactivation of the flax rust resistance gene *M* associated with loss of a repeated unit within the leucine-rich repeat coding region. Plant Cell **9,** 641-651.

Arumuganathan, E., and Earle, E.D. (1991). Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. **9,** 208-218.

Bai, J., Pennill, L.A., Ning, J., Lee, S.W., Ramalingam, J., Webb, C.A., Zhao, B., Sun, Q., Nelson, J.C., Leach, J.E., and Hulbert, S.H. (2002). Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. Genome Res. **12,** 1871-1884.

Bailey, T.L., and Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. Proc. Int. Conf. Intell. Syst. Mol. Biol. **3,** 21-29.

Bailey, T.L., and Gribskov, M. (1998). Methods and statistics for combining motif match scores. J. Comput. Biol. **5,** 211-221.

Baumgarten, A., Cannon, S., Spangler, R., and May, G. (2003). Genome-level evolution of resistance genes in *Arabidopsis thaliana*. Genetics **165,** 309-319.

Bendahmane, A., Kanyuka, K., and Baulcombe, D.C. (1999). The *Rx* gene from potato controls separate virus resistance and cell death responses. Plant Cell **11,** 781-792.

Bent, A.F., Kunkel, B.N., Dahlbeck, D., Brown, K.L., Schmidt, R., Giraudat, J., Leung, J., and Staskawicz, B.J. (1994). *RPS2* of *Arabidopsis thaliana*: A leucine-rich repeat class of plant disease resistance genes. Science **265,** 1856-1860.

Bhattramakki, D., Dong, J., Chhabra, A.K., and Hart, G. (2000). An integrated SSR and RFLP linkage map of *Sorghum bicolor* (L.) Moench. Genome **43,** 988-1002.

Bittner-Eddy, P.D., Crutem I.R., Holub, E.B., and Beynon, J.L. (2000). *RPP13* is a simple locus in *Arabidopsis thaliana* for alleles that specify downy mildew resistance to different avirulence determinants in *Peronospora parasitica*. Plant J. **21,** 177-188.

Botella, M.A., Parker, J.E., Frost, L.N., Bittner-Eddy, P.D., Beynon, J.L., Daniels, M.J., Holub, E.B., and Jones, J.D. (1998). Three genes of the *Arabidopsis RPP1* complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. Plant Cell **10,** 1847-1860.

Bowers, J.E., Abbey, C., Anderson, S., Changg, C., Draye, X., Hoppe, A.H., Jessup, R., Lemke, C., Lennington, J., Li, Z., Lin, Y.-R., Liu, S.-C., Luo, L., Marler, B.S., Ming, R., Mitchell, S.E., Qiang, D., Reischmann, K., Schulze, S.R., Skinner, D.N., Wang, Y.-W., Kresovich, S., Schertz, K.F., and Paterson, A.H. (2003). A high-density genetic recombination map of sequence-tagged sites for Sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics **165,** 367-386.

Brommonschenkel, S.H., Frary, A., Frary, A., and Tanksley, S.D. (2000). The broad-spectrum tospovirus resistance gene *Sw-5* of tomato is a homolog of the root-knot nematode resistance gene *Mi*. Mol. Plant-Microbe Interact. **13,** 1130-1138.

Bryan, G.T., Wu, K.-S., Farrall, L., Jia, Y., and Hershey, H.P. (2000). A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene *Pi-ta*. Plant Cell **12,** 2033-2046.

Buschges, R., Holricher, K., Panstruga, R., Simons, G., and Wolter, M., Frijters, A., Van Daelen, R., Van der Lee, T., Diergaarde, P., Groenendijk, J., Topsch, S., Vos, P., Salamini, F., and Schulze-Lefert, P. (1997). The barley *Mlo* gene: A novel control element of plant pathogen resistance. Cell **88,** 695-705.

Cai, D., Kleine, M., Kifle, S., Harloff, H.J., and Sandal, N.N., Marcker, K.A., Klein-Lankhorst, P.M., Salentijn, M.J., Lange, W., Steikema, W.J., Wyss, U., Grundler, M.W., and Jung, Christian. (1997). Positional cloning of a gene for nematode resistance in sugar beet. Science **275,** 832-834.

Cannon, S.B., Zhu, H., Baumgarten, A.M., Spangler, R., May, G., Cook, D.R., and Young, N.D. (2002). Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies. J. Mol. Evol. **54,** 548-562.

Collins, N., Drake, J., Ayliffe, M., Sun, Q., and Ellis, J., Hulbert, S., and Pryor, T. (1999). Molecular characterization of the maize *Rp1-D* rust resistance haplotype and its mutants. Plant Cell **11,** 1365-1376.

Cooley, M.B., Pathirana, S., Wu, H.J., Kachroo, P., and Klessig, D.F. (2000). Members of the *Arabidopsis HRT/RPP8* family of resistance genes confer resistance to both viral and oomycete pathogens. Plant Cell **12,** 663-676.

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1979). Atlas of Protein Sequence and Structure, vol. 5. (Washington, D.C.: National Biomedical Research Foundation).

Deslandes, L., Olivier, J., Theulieres, F., Hirsch, J., Feng, D.X., Bittner-Eddy, P., Beynon, J., and Marco, Y. (2002). Resistance to *Ralstonia solanacearum* in *Arabidopsis thaliana* is conferred by the recessive *RRS1-R* gene, a member of a novel family of resistance genes. Proc. Natl. Acad. Sci. USA **99,** 2404-2409.

Dixon, M.S., Hatzixanthis, K., Jones, D.A., Harrison, K., and Jones, J.D. (1998). The tomato *Cf-5* disease resistance gene and six homologs show pronounced allelic variation in leucine-rich repeat copy number. Plant Cell **10,** 1915-1925.

Dje, Y., Heuertz, M., Lefebvre, C., and Vekemans, X. (2000). Assessment of genetic diversity within and among germplasm accessions in cultivated sorghum using microsatellite markers. Theor. Appl. Genet. **100,** 918-925.

Dodds, P.N., Lawrence, G.J., and Ellis, J.G. (2001). Six amino acid changes confined to the leucine-rich repeat β-strand/β-turn motif determine the difference between the *P* and *P2* rust resistance specificities in flax. Plant Cell **13,** 163-178.

Doggett, H. (1988). Sorghum, edn 2. (New York: John Wiley).

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics **14,** 755-763.

Ellis, J., and Jones, D. (1998). Structure and function of proteins controlling strain-specific pathogen resistance in plants. Curr. Opin. Plant Biol. **1,** 288-293.

Ellis, J.G., Lawrence, G.J., Luck, J.E., and Dodds, P.N. (1999). Identification of regions in alleles of the flax rust resistance gene *L* that determine differences in gene-for-gene specificity. Plant Cell **11,** 495-506.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39,** 783-791.

Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. Syst. Zool. **19,** 99-113.

Flor, H.H. (1971). Current status of the gene-for-gene concept. Annu. Rev. Phytopathol. **9,** 275-296.

Gassmann, W., Hinsch, M.E., and Staskawicz, B.J. (1999). The *Arabidopsis RPS4* bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. Plant J. **20,** 265-277.

Goff, S.A., Ricke, D., Lan, T.-H., Presting, G., Wang, R. et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296,** 92-100.

Grant, M., Godiard, L., Straube, E., Ashfield, T., Leward, J., Sattler, A., Innes, R., and Dangl, J. (1995). Structure of the *Arabidopsis RPM1* gene enabling dual specificity disease resistance. Science **269,** 843-846.

Haldane, J.B.S. (1919). The combination of linkage values, and the calculation of distance between the loci of linked factors. J. Genet. **8,** 299-309.

Hammond-Kosack, K.E., and Jones, J.D. (1996). Resistance gene-dependent plant defense responses. Plant Cell **8,** 1773-1791.

Hammond-Kosack, K.E., and Jones, J.D. (1997). Plant disease resistance genes. Annu. Rev. Plant Physiol. Plant Mol. Biol. **48,** 575-607

Helentjaris, T., Slocum, M., Wright, S., Schaefer, A., and Nienhuis, J. (1986). Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphisms. Theor. Appl. Genet. **72,** 761-769.

Hickey, D.A., Bally-Cuif, L., Abukashawa, S., Payant, V., and Benkel, B.F. (1991). Concerted evolution of duplicated protein-coding genes in *Drosophila*. Proc. Natl. Acad. Sci. USA **88,** 1611-1615.

Holub, E. (2001). Arms race is ancient history in *Arabidopsis*, the wildflower. Nat. Rev. Genet. **2,** 516-527.

Hulbert, S.H., Webb, C.A., Smith, S.M., and Sun, Q. (2001). Resistance gene complexes: Evolution and utilization. Annu. Rev. Phytopathol. **39,** 285-312.

Irigoyen, M.L., Loarce, Y., Fominaya, A., and Ferrer, E. (2004). Isolation and mapping of resistance gene analogs from the *Avena strigosa* genome. Theor. Appl. Genet. **109,** 713-724.

Johal, G.S., and Briggs, S.P. (1992). Reductase activity encoded by the *HM1* disease resistance gene in maize. Science **258,** 985-987.

Jones, D.A., and Jones, J.D.G. (1997). The role of leucine-rich repeat proteins in plant defenses. Adv. Bot. Res. **24,** 90-167.

Jones, D.A., Thomas, C.M., Hammond-Kosack, K.E., Balint-Kurti, P.J., and Jones, J.D.G. (1994). Isolation of the tomato *Cf-9* gene for resistance to *Cladosporium fulvum* by transposon tagging. Science **266,** 789-793.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences (CABIOS) **8,** 275-282.

Kanazin, V., Marek, L.F., and Shoemaker, R.C. (1996). Resistance gene analogs are conserved and clustered in soybean. Proc. Natl. Acad. Sci. USA **93,** 11746-11750.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution. (Cambridge: Cambridge University Press).

Klein, R.R., Cartinhour, S.W., Ulanch, P.E., Dong, J., Obert, J.A., Morishige, D.T., Schlueter, S.D., Childs, K.L., Ale, M., Mullet, J.E., and Klein, P.E. (2000). A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. Genome Res. **10,** 789-807.

Kong, L., Dong, J., and Hart, G.E. (2000). Isolation, characterization, and linkage mapping of *Sorghum bicolor* (L.) Moench DNA simple-sequence-repeats (SSRs). Theo. Appl. Genet. **101,** 438-448.

Landschulz, W.H., Johnson, P.F., and McKnight, S.L. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. Science **240,** 1759-1762.

Lawrence, G.J., Finnegan, E.J., Ayliffe, M.A., and Ellis, J.G. (1995). The *L6* gene for flax rust resistance is related to the *Arabidopsis* bacterial resistance gene *RPS2* and the tobacco viral resistance gene *N*. Plant Cell **7,** 1195-1206.

Leister, D., Kurth, J., Laurie, D.A., Yano, M., Sasaki, T., Devos, K., Graner, A., and Schulze-Lefert, P. (1998). Rapid reorganization of resistance gene homologues in cereal genomes. Proc. Natl. Acad. Sci. USA **95,** 370-375.

Li, P., Nijhawan, D., Budihardjo, I., Srinvasula, S.M., Ahmad, M., Alnemri, E.S., and Wang, X. (1997). Cytochrome c and dATP-dependent formation of *Apaf*-1/caspase-9 complex initiates an apoptotic protease cascade. Cell **91,** 479-487.

Lynch, E., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. Genetics **154,** 459-473.

Madsen, L.H., Collins, N.C., Rakwalska, M., Backes, G., Sandal, N., Krusell, L., Jensen, J., Waterman, E.H., Jahoor, A., Ayliffe, M., Pryor, A.J., Langridge, P., Schulze-Lefert, P., and Stougaard, J. (2003). Barley disease resistance gene analogs of the NBS-LRR class: Identification and mapping. Mol. Gen. Genomics **269,** 150-161.

Maniatis, F., Fritsch, E.F., and Sambrook, J. (1982). Molecular Cloning: A Laboratory Manual. (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

Martin, G.B., Brommonschenkel, S.H., Chunwongse, J., Frary, A., and Ganal, M.W., Spivey, R., Wu, T., Earle, E.D., and Tanksley, S.D. (1993). Map-based cloning of a protein kinase gene conferring disease resistance in tomato. Science **262,** 1432-1436.

McDowell, J.M., Dhandaydham, M., Long, T.A., Aarts, M.G., and Goff, S., Holub, E.B., and Dangl, J.L. (1998). Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the *RPP8* locus of *Arabidopsis*. Plant Cell **10,** 1861-1874.

McIntyre, C.L., Hermann, S.M., Casu, R.E., Knight, D., Drenth, J., Tao, Y., Brumbley, S.M., Godwin, I.D., Williams, S., Smith, G.R., and Manners, J.M. (2004). Homologues of the maize rust resistance gene *Rp1-D* are genetically associated with a major rust resistance QTL in sorghum. Theor. Appl. Genet. **109,** 875-883.

Medzhitov, R., Preston-Hurlburt, P., and Janeway, C.A.J. (1997). A human homologue of the *Drosophila* Toll protein signals activation of adaptive immunity. Nature **388,** 394-397.

Menz, M.A., Klein, R.R., Mullet, J.E., Obert, J.A., Unruh, N.C., and Klein, P.E. (2002). A high-density genetic map of *Sorghum bicolor* (L.) Moench based on 2926 AFLP, RFLP and SSR markers. Plant Mol. Biol. **48,** 483-499.

Meyers, B.C., Chin, D.B., Shen, K.A., Sivaramakrishnan, S., Lavelle, D.O., Zhang, Z., and Michelmore, R.W. (1998). The major resistance gene cluster in lettuce is highly duplicated and spans several megabases. Plant Cell **10,** 1817-1832.

Meyers, B.C., Dickerman, A.W., Michelmore, R.W., Sivaramakrishnan, S., Sobral, B.W., and Young, N.D. (1999). Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. Plant J. **20,** 317-332.

Meyers, B.C., Kozik, A., Griego, A., Kuang, H., and Michelmore, R.W. (2003). Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. Plant Cell **15,** 809-834.

Meyers, B.C., Morgante, M., and Michelmore, R.W. (2002). TIR-X and TIR-NBS proteins: Two new families related to disease resistance TIR-NBS-LRR proteins encoded in *Arabidopsis* and other plant genomes. Plant J. **32,** 77-92.

Michelmore, R. (2000). Genomic approaches to plant disease resistance. Curr. Opin. Plant Biol. **3,** 125-131.

Michelmore, R.W., and Meyers, B.C. (1998). Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res. **8,** 1113-1130.

Milligan, S.B., Bodeau, J., Yaghoobi, J., Kaloshian, I., Zabel, P., and Williamson, V.M. (1998). The root knot nematode resistance gene *Mi* from tomato is a member of the leucine zipper, nucleotide binding, leucine-rich repeat family of plant genes. Plant Cell **10,** 1307-1319.

Mindrinos, M., Katagiri, F., Yu, G.-L., and Ausubel, F.M. (1994). The *A. thaliana* disease resistance gene *RPS2* encodes a protein containing a nucleotide-binding site and leucine-rich repeats. Cell **78,** 1089-1099.

Mondragon-Palomino, M., Meyers, B.C., Michelmore, R.W., and Gaut, B.S. (2002). Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. Genome Res. **12,** 1305-1315.

Murray, M.G., and Thompson, W.F. (1980). Rapid isolation of high-molecular-weight plant DNA. Nucleic Acids Res. **8,** 4321-4325.

Noir, S., Combes, M.-C., Anthony, F., and Lashermes, P. (2001). Origin, diversity and evolution of NBS-type disease-resistance gene homologues in coffee trees (*Coffea* L.). Mol. Genet. Genomics **265,** 654-662.

Ori, N., Eshed, Y., Paran, I., Presting, G., and Aviv, D., Tanksley, S., Zamir, D., and Fluhr, R. (1997). The *I2C* family from the wilt disease resistance locus *I2* belongs to the nucleotide binding, leucine-rich repeat superfamily of plant resistance genes. Plant Cell **9,** 521-532.

Otto, S., and Young, P. (2002). The evolution of gene duplicates. Adv. Genet. **46,** 451-483.

Pan, Q., Wendel, J., and Fluhr, R. (2000). Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. J. Mol. Evol. **50,** 203-213.

Parker, J.E., Coleman, M.J., Szabo, V., Frost, L.N., and Schmidt, R., Van der Biezen, E.A., Moores, T., Dean, C., Daniels, M.J., and Jones, J.D.G. (1997). The *Arabidopsis* downy mildew resistance gene *Rpp5* shares similarity to the toll and interleukin-1 receptors with *N* and *L6*. Plant Cell **9,** 879-894.

Parniske, M., Hammond-Kosack, K.E., Golstein, C., Thomas, C.M., Jones, D.A., Harrison, K., Wulff, B.B.H., and Jones, J.D.G. (1997). Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the *Cf-4/9* locus of tomato. Cell **91,** 821-832.

Peng, Y., Schertz, K.F., Cartinhour, S., and Hart, G.E. (1999). Comparative genome mapping of *Sorghum bicolor* (L.) Moench using an RFLP map constructed in a population of recombinant inbred lines. Plant Breed. **118,** 225-235.

Peterson, D.G., Schulze, S.R., Sciara, E.B., Lee, S.A., Bowers, J.E., Nagel, A., Jiang, N., Tibbitts, D.C., Wessler, S.R., and Paterson, A.H. (2002). Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res. **12,** 795-807.

Quint, M., Mihaljevic, R., Dussle, C.M., Xu, M.L., Melchinger, A.E., and Lubberstedt, T. (2002). Development of RGA-CAPS markers and genetic mapping of candidate genes for sugarcane mosaic virus resistance in maize. Theor. Appl. Genet. **105,** 355-363.

Ramakrishna, W., Emberton, J., SanMiguel, P., Ogden, M., Llaca, V., Messing, J., and Bennetzen, J.F. (2002). Comparative sequence analysis of the sorghum *Rph* region and the Maize *Rp1* resistance gene complex. Plant Physiol. **130,** 1728-1738.

Ramalingam, J., Vera Cruz, C.M., Kukreja, K., Chittoor, J.M., Wu, J.-L., Lee, S.W., Baraoidan, M., George, M.L., Cohen, M.B., Hulbert, S.H., Leach, J.E., and Leung, H. (2003). Candidate defense genes from rice, barley, and maize and their association with qualitative and quantitative resistance in rice. Mol. Plant-Microbe Interact. **16,** 14-24.

Reed, K.C., and Mann, D.A. (1985). Rapid transfer of DNA agarose gels to nylon membranes. Nucleic Acids Res. **13,** 7207-7221.

Richly, E., Kurth, J., and Leister, D. (2002). Mode of amplification and reorganization of resistance genes during recent *Arabidopsis thaliana* evolution. Mol. Biol. Evol. **19,** 76-84.

Richter, T.E., and Ronald, P.C. (2000). The evolution of disease resistance genes in plants. Plant Mol. Biol. **3,** 157-161.

Rossi, M., Goggin, F.L., Milligan, S.B., Kaloshian, I., Ullman, D.E., and Williamson, V.M. (1998). The nematode resistance gene *Mi* of tomato confers resistance against the potato aphid. Proc. Natl. Acad. Sci. USA **95,** 9750-9754.

Saghai-Maroof, M.A., Soliman, K.M., Jorgensen, R.A., and Allard, R.W. (1984). Ribosomal DNA spacer-length polymorphism in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. USA **81,** 8014-8018.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol **4,** 406-425.

Salmeron, J.M., Oldroyd, G.E.D., Tommens, C.M.T., Scofield, S.R., Kim, H.-S., Lavelle, D.T., Dahlbeck, D., and Staskawicz, B.J. (1996). Tomato *Prf* is a member of the leucine-rich repeat class of plant disease resistance genes and lies embedded within the *Pto* kinase gene cluster. Cell **86,** 123-133.

Saraste, M. Sibbald, P.R., and Wittinghofer, A. (1990). The P-loop: A common motif in ATP- and GTP-binding proteins. Trends Biochem. Sci. **15,** 430-434.

Sawant, S.V., Kiran, K., Singh, P.K., and Tuli, R. (2001). Sequence architecture downstream of the initiator codon enhances gene expression and protein stability in plants. Plant Physiol. **126,** 1630-1636.

Shen, K.A., Meyers, B.C., Islam-Faridi, M.N., Chin, D.B., Stelly, D.M., and Michelmore, R.W. (1998). Resistance gene candidates identified by PCR with degenerate oligonucleotide primers map to clusters of resistance genes in lettuce. Mol. Plant-Microbe Interact. **11,** 815-823.

Simons, G., Groenendijk, J., Wijbrandi, J., Reijans, and M., Groenen, J., Diergaarde, P., Van der Lee, T., Bleeker, M., Onstenk, J., de Both, M., Haring, M., Mes, J., Cornelissen, B., Zabeau, M., and Vos, P. (1998). Dissection of the *Fusarium I2* gene cluster in tomato reveals six homologs and one active gene copy. Plant Cell **10,** 1055-1068.

Song, R., Llaca, V., and Messing, J. (2002). Mosaic organization of orthologous sequences in grass genomes. Genome Res. **12,** 1549-1555.

Song, W.-Y., Wang, G.-L., Chen, L.-L., Kim, H.-S., and Pi, L.-Y., Holsten, T., Gardner, J., Wang, B., Zhai, W.-X., Zhu, L.-H., Fauquet, C., and Ronald, P. (1995). A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. Science **270,** 1804-1806.

Srinivasula, S.M., Ahmad, M., Fernandes-Alnemri, T., and Alnemri, E.S. (1998). Autoactivation of procaspase-9 by Apaf-1-mediated oligomerization. Mol. Cell **1,** 949-957.

Staskawicz, B.J., Ausubel, F.M. Baker, B.J., Ellis, J.G., and Jones, J.D.G. (1995). Molecular genetics of plant disease resistance. Science **268,** 661-667.

Tai, T.H., Dahlbeck, D., Clark, E.T., Gajiwala, P., Pasion, R., Whalen, M.C., Stall, R.E., and Staskawicz, B.J. (1999). Expression of the *Bs2* pepper gene confers resistance to bacterial spot disease in tomato. Proc. Natl. Acad. Sci. USA **96,** 14153-14158.

Takken, F.L., Thomas, C.M., Joosten, M.H., Golstein, C., and Westerink, N., Hille, J., Nijkamp, H.J.J., De Wit, P.J.G.M., and Jones, J.D.G. (2000). A second gene at the tomato *Cf-4* locus confers resistance to *Cladosporium fulvum* through recognition of a novel avirulence determinant. Plant J. **20,** 279-288.

Tameling, W.L., Elzinga, S.D.J., Darmin, P.S., Vossen, J.H., Takken, F.L.W., Harling, M.A., and Cornelissen, B. (2002). The tomato *R* gene products I-2 and Mi-1 are functional ATP binding proteins with ATPase activity. Plant Cell **14,** 2929-2939.

Tao, Q., and Zhang, H.-B. (1998). Cloning and stable maintenance of DNA fragments over 300 kb in *Escherichia coli* with conventional plasmid-based vectors. Nucleic Acids Res. **26,** 4901-4909.

Tao, Y.Z., Jordan, D.R., Henzell, R.G., and McIntyre, C.L. (1998). Identification of genomic regions for rust resistance in sorghum. Euphytica **103,** 287-292.

Thomas, C.M., Jones, D.A., Parniske, M., Harrison, K., and Balint-Kurti, P.J., Hatzixanthis, K., and Jones, J.D.G. (1997). Characterization of the tomato *Cf-4* gene for resistance to *Cladosporium fulvum* identifies sequences that determine recognitional specificity in *Cf-4* and *Cf-9*. Plant Cell **9,** 2209-2224.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment trough sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22,** 4673-4680.

Traut, T.W. (1994). The functions and consensus motifs of nine types of peptide segments that form different types of nucleotide-binding sites. Eur. J. Biochem. **222,** 9-19.

Van der Biezen, E.A., and Jones, J.D.G. (1998). The NB-ARC domain: A novel signaling motif shared by plant resistance gene products and regulators of cell death in animals. Curr. Biol. **8,** 226-227.

Van der Hoorn, R.A.L., De Wit, P.J.G.M., and Joosten, M.H.A.J. (2002). Balancing selection favors guarding resistance proteins. Trends in Plant Science **7,** 67-71.

Van der Vossen, E.A.G., Rouppe van der Voort, J.N.A.M., Kanyuka, K., Bendahmane, A., Sandbrink, H., Baulcombe, D.C., Bakker, J., Stiekema, W.J., and Klein-Lankhorst, R.M. (2000). Homologues of a single resistance-gene cluster in potato confer resistance to distinct pathogens: A virus and a nematode. Plant J. **23,** 567-576.

Vos, P., Simons, G., Jesse, T., Wijbrandi, J., and Heinen, L., Hogers, R., Frijters, A., Groenendijk, J., Diergaarde, P., Reijans, M., Fierens-Onstenk, J., de Both, M., Peleman, J., Liharska, T., Hontelez, J., and Zabeau, M. (1998). The tomato *Mi-1* gene confers resistance to both root-knot nematodes and potato aphids. Nat. Biotechnol. **16,** 1365-1369.

Walsh, J.B. (1987). Sequence-dependent gene conversion: Can duplicated genes diverge fast enough to escape conversion? Genetics **117,** 543-557.

Walsh, J.B. (1995). How often do duplicated genes evolve new functions? Genetics **139,** 421-428.

Wang, Z.X., Yano, M., Yamanouchi, U., Iwamoto, M., and Monna, L., et al. (1999). The *Pib* gene for rice blast resistance belongs to the nucleotide binding and leucine-rich repeat class of plant disease resistance genes. Plant J. **19,** 55-64.

Warren, R.F., Henk, A., Mowery, P., Holub, E., and Innes, R.W. (1998). A mutation within the leucine-rich repeat domain of the *Arabidopsis* disease resistance gene *RPS5* partially suppresses multiple bacterial and downy mildew resistance genes. Plant Cell **10,** 1439-1452.

Wei, F., Gobelman-Werner, K., Morroll, S.M., Kurth, J., Mao, L., Wing, R., Leister, D., Schulze-Lefert, P., and Wise, R.P. (1999). The *Mla* (powdery mildew) resistance cluster is associated with three NBS-LRR gene families and suppressed recombination within a 240-kb DNA interval on chromosome 5S (1HS) of barley. Genetics **153,** 1929-1948.

Whitham, S., McCormick, S., and Baker, B. (1996). The *N* gene of tobacco confers resistance to tobacco mosaic virus in transgenic tomato. Proc. Natl. Acad. Sci. USA **93,** 8776-8781.

Wilson, W.A., Harrington, S.E., Woodman, W.L., Lee, M., Sorrells, M.E., and McCouch, S.R. (1999). Inferences on the genome structure of progenitor maize through comparative analysis of rice, maize and the domesticated panicoids. Genetics **153,** 453-473.

Woo, S.-S., Jiang, J., Gill, B.S., Paterson, A.H., and Wing, R.A. (1994). Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*. Nucleic Acids Res. **22,** 4922-4931.

Xiao, S., Ellwood, S., Calis, O., Patrick, E., and Li, T., Coleman, M., and Turner, J.G. (2001). Broad-spectrum mildew resistance in *Arabidopsis thaliana* mediated by *Rpw8*. Science **291,** 118-120.

Yang, R.B., Mark, M.R., Gram, A., Huang, A., Xie, M.H., Zhang, M., Goddard, A. Wood, W.I., Gurney, A.L., and Godowski, P.J. (1998). Toll-like receptor-2 mediates lipopolysaccharide-induced cellular signaling. Nature **395,** 284-288.

Yoshimura, S., Yamanouchi, U., Katayose, Y., Toki, S., Wang, Z.-X., Kono, I., Kurata, N., Yano, M., Iwata, N., and Sasaki, T. (1998). Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. Proc. Natl. Acad. Sci. USA **95,** 1663-1668.

Young, N.D. (2000). The genetic architecture of resistance. Curr. Opin. Plant Biol. **3,** 285-290.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S. et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296,** 79-92.

Yu, Y.G., Buss, G.R., and Maroof, M.A. (1996). Isolation of a superfamily of candidate disease-resistance genes in soybean based on a conserved nucleotide-binding site. Proc. Natl. Acad. Sci. USA **93,** 11751-11756.

Zhou, F., Kurth, J., Wei, F., Elliott, C., Vale, G., Yahiaoui, N., Keller, B., Somerville, S., Wise, R., and Schulze-Lefert, P. (2001). Cell-autonomous expression of barley *Mla1* confers race-specific resistance to the powdery mildew fungus via a *Rar1* independent signaling pathway. Plant Cell **13,** 337-350.

Zhou, T., Wang, Y., Chen, J.-Q., Araki, H., Jing, Z., Jiang, K., Shen, J., and Tian, D. (2004). Genome-wide identification of NBS genes in *japonica* rice reveals significant expansion of divergent non-TIR NBS-LRR genes. Mol. Gen. Genomics **271,** 402-415.

Zhu, H., Cannon, S.B., Young, N.D., and Cook, D.R. (2002). Phylogeny and genomic organization of the TIR and non-TIR NBS-LRR resistance gene family in *Medicago truncatula*. Mol. Plant-Microbe Interact. **15,** 529-539.

APPENDIX A

A-1

ALIGNMENT OF CONSERVED MOTIF P-LOOP IDENTIFIED BY MEME

NAME	P-VALUE	SITES
TC81885	8.19e-23	GGRSSAEQWI **VSIVGFGGLGKTTLAKAVYDK** IKPQFDCTAF
CD212839	2.87e-22	ERLRV **VSIVGFGGLGKTTLANEVYRD** LRDNLSGNPE
BG412236	3.31e-22	VDGEPQQLRV **ISIVGFGGIGKTTLARAVYDS** PQAKEKFQCR
BM327689	4.44e-22	VDGEPQQLRV **ISIVGFGGLGKTTLARAVYDS** PHAKETFHCR
RHOH113F05G1	1.35e-21	DEASSTRYSS **LAIVGAGGMGKSTLAQYVYND** ERIKEGFDVR
AAM94294	3.83e-21	NGVPFQQGKV **VSIVGFGGLGKTTLAKVVYEK** IRSLFHCCAF
AAM94297	5.57e-21	QEASKQHGRV **VSIVGCGGLGKTTLANVVYQK** IRTQFDCWAF
BZ341506	1.02e-20	HEASKMDLLV **LPIVGMGGLGKTTFIQLVYND** PAIQKHFQLQ
BZ343608	2.32e-20	PGISKSGPRV **VSVVGMGGLGKTTLTKKVYDS** KDLGDIFEIR
NP853482	2.60e-20	DEASSTRYSS **LAIIGAGGMGKSTLAQYVYND** KRIEEGFDIR
BZ369917	4.07e-20	KSKSNNVVVA **VAITGMGGIGKTTLARMVFND** NKIEENFEDR
AAM94295	4.55e-20	NEVPIQKGKI **VTIVGFGGLGKTTLAHAVFDK** IRPGFDCCAS
OX158E07B1	1.34e-19	DAPDKKITKK **VSIVGVGGLGKTTVAKAVYEN** LKSQFDCAAF
TC85900	1.49e-19	DEASSTRYSS **LAIIGAGGMGKSTLVQYVYND** KRIEEGFDIR
BZ367728	4.62e-19	SKEDDEQTMV **ISVWGMGGLGKTTLVKEVYQS** QELSDLFEKR
AAD27570	4.62e-19	VDDGAQGVKV **VSIVGCGGLGKTTIANQVYIN** IAEKFDCQAF
BZ350423	5.65e-19	ESNEGENVWI **VSIVGLGGSGKTTLAKQICHD** VKIKQHFKST
BM326535	7.60e-19	GKHSTEILTV **IPIVGPGGIRKTTLAQHIYHS** PDVQDHFDVR
TC76961	1.02e-18	HESSNTGPRV **VSLVGMGGIGKTTLTKKVFDS** NDLSDKFGTR
TC80519	6.70e-18	LREGKKKVDV **FAIVGAVGIGKTTLAREIYND** DRMTENFPIC
BZ350669	1.38e-17	DHANNDDLLV **LPIVGLGGLGKPTFVQLVYSD** PEIEKHFQFL
BZ334356	1.64e-17	SSSSRQQSNI **ISIVGFGGLGKTTLANSLLQD** LKSKFDCHIF
TC89319	3.59e-17	VTQSGKTLSV **LPIVGPGGIGKTTFTQHLVNH** TRIKQCFHDI
BZ345488	3.91e-17	DSDEGLNGWI **VSIIGIGGSGKTTLAKLICLD** KRTKEHFKDS
CD209645	5.05e-17	GTSSSKCCSV **ICIHGIAGSGKTTLAQYVCDH** ENENREKYFN
BH246154	2.64e-16	TEDMEPRRTL **VAVWGMGGVGKTTLVTNVFRE** VAASFHFDCA
TC79945	4.93e-15	SSDQNHKVQV **LPIVGEACIGKTTVAQLVITD** ERILLHFKLR
TC89312	1.18e-14	DLLEKGESNI **IGVWGQGGIGKTTLLHAFNND** LEKKDHNYQV
BZ348590	2.40e-14	PDGSERMFRA **AGIAGIHGSGKTALAQKVFVH** DKAKDNFALR
BM325897	7.29e-14	THDDCPNLGI **LPIIGPHRVGKKTLVQHACKD** ERVRGFFSKI
BZ423246	2.42e-12	SRSTRAAITV **LPIIGGCRVGKKTLVGNICSD** DRIRSCYPCI

P-loop motif consensus sequence **VSIVGFGGLGKTTLAQxVYND**

A-2

ALIGNMENT OF CONSERVED MOTIF RNBS-A IDENTIFIED BY MEME

NAME	P-VALUE	SITES
AAM94297	3.36e-21	NVVYQKIRTQ **FDCWAFVSVSQTPDMRRLFEGILSELGKD** INEETRDVRH
AAM94295	3.90e-21	HAVFDKIRPG **FDCCASVSVSQTPDLKKLLKGILYQLDKK** YEDINEKPLD
BZ343608	4.52e-21	VYDSKDLGDI **FEIRAWIAVSQSFDPKELLKEMIKQLFGA** HSLKEFLEEH
NP239124	5.24e-21	VFHSIDIVGN **FSSRAWITVSQSFDKKELLKELIKQLFGD** GSSKEHSRGL
AAM94294	1.44e-20	KVVYEKIRSL **FHCCAFISVSQTPDLKKLFKELLYDLDKN** INAETLDERR
TC85900	7.61e-20	VYNDKRIEEG **FDIRMWVCISRKLDVRRHTREIIESATNG** ECPCIDNLDT
NP853482	7.61e-20	VYNDKRIEEG **FDIRMWVCISRKLDVRRHTREIIESATNG** ECPCIDNLDT
Sb_RGA55	9.97e-20	VYKNQNITRT **FNCHAWVTVSQTYQVEELLREIINQLIDQ** RASMASGFMT
RHOH113F05G1	3.71e-19	VYNDERIKEG **FDVRIWVCISRKLDVRRHTRKIIESATNG** ECPCIGNLDT
TC80927	1.16e-18	TREKITDQ **FSCAAFVSVSQKPNMISLLWELLSQIGSH** GGDLGLMAIG
BZ334356	1.68e-18	NSLLQDLKSK **FDCHIFVSVSVNPDIKKIFKNILLQLDEN** EYSRIDEGWE
Sb_RGA50	3.10e-18	FNDGETIEKQ **FEVRLWVHVSQEFDFEKLIKKLFEAFADK** DPGQPSLPYM
AAD27570	3.50e-18	NQVYINIAEK **FDCQAFVSLTQNPDMVIIFQSILTQVKKD** ECDSTSSCDK
TC75876	5.01e-18	ARGKLKAQ **FECEAFVSVSLDPRMDQVFKSMLRQLDKD** KYNNIKGEMW
TC81885	6.36e-18	KAVYDKIKPQ **FDCTAFISVFRDPDIIKIFKDMLYELDNK** EYWDIHNIAL
TC76961	7.15e-18	VFDSNDLSDK **FGTRAWITVSQSFEQKEIFKEMVKHLFGA** ESLHKLLEDH
Sb_RGA130	1.62e-17	QHINEDMKSH **FHVRVWVCISQNFSASRLAQEIAKQIPKL** DNEKENESAE
TC87218	2.03e-17	VYNDARIEAR **FGMRAWVCVWDRSDEVELTREILQSIGCA** DDAPCDDGLS
TC89319	2.28e-17	VNHTRIKQCF **HDINIWICVSTNFDVLKLTKEMLSCLPAT** ENEENNETTT
BH246154	2.28e-17	VFREVAASFH **FDCAAWVSVSKNFTREDLLKRVLKELQRD** VSAGVPKDVE
Sb_RGA80	3.20e-17	VYNDAVVQDH **FNKRIWISVSIHFDEVRLTREMLDCLSDG** VSKHDEIINL
NP239122	5.01e-17	VYNHEKIKGT **FSMQAWICVSKEYSEDALLKEVLRNIGID** YKQDETTGEL
BZ340437	7.80e-17	V **FGLSIWVWVSNNFDAATVIRMILESIDKK** NPTVDVLEIL
TC79945	1.08e-16	VITDERILLH **FKLRPWVHVSNEFNIRRITADIIESIEGS** SPRFN
BZ345488	1.08e-16	CLDKRTKEHF **KDSILWVHVSQEFDLEKLIGKLFESIAKK** KADRHTQQYM
Sb_RGA75	1.34e-16	HIYNKEAETY **FDVRIWACVSTDFSVPRLLKDILESKSLH** ELSKGLTGTP
OX158E07B1	2.06e-16	KAVYENLKSQ **FDCAAFVSVGRDLDLVKVFKDILFDLDKE** EYKDIHETKR
Sb_RGA181	2.30e-16	IXXXDKIKGS **FSXQAWICVSQQYSDISVLKEVLRNIGVD** YKHDETVGEL
CD212839	7.25e-16	DNLSGNPEKS **FSCKAIISVSQRPDMVNLLKSLFTKVSGQ** TADHTYDLPG
BM326535	1.09e-15	IYHSPDVQDH **FDVRVWTCVSLNFNVNKLIEEIQGYIPKI** D
Sb_RGA125	5.88e-15	TNVYEREKIN **FSATAWMVVSQTYTIEALLRKLLMKVGGE** QQVPPNIDKL
BZ350423	9.50e-15	CHDVKIKQHF **KSTIFWVHVSEEFDVKELIGKLFETILEQ** KSDLHAQQHM
Sb_RGA182	7.27e-14	KLLFKDIQFN **KYSRVWVYVSEIFNLKKIGNSIISQVSKT** ESQITMQMIH

RNBS-A motif consensus sequence **FDCRAWVSVSQxFDVKKLLKEILEQLxKD**

A-3

ALIGNMENT OF CONSERVED MOTIF KINASE-2 IDENTIFIED BY MEME

NAME	P-VALUE	SITES
TC76961	1.10e-18	DYLSKRLKET **RYLIVLDDVWTIDAW** NRIKVTFQDS
Sb_RGA55	7.54e-17	EVIQSYLLDK **KYLIVLDDVWDKDAW** LFLNHAFVRN
AAM94306	3.64e-16	SEVREFLEKK **RYLIVIDDIWDITAW** KMIKCALPDN
AAM94295	4.75e-16	NELRKFLRRK **RYFIVIDDIWDISVW** RMIKCALPHS
AAM94294	1.28e-15	NVLREFLIPK **RYLVVIDDIWDVSVW** EVIKCALPEN
TC80927	9.13e-15	DRLRSDLENQ **RYLVVIDDVWTKSPW** EIIQCALPNN
AAM94297	1.08e-14	DAIGKFLQTK **RYCIVIDDIWDISVW** KMIRCALPDN
BZ343608	2.00e-14	NYLRGRLLER **KYLVVLDDVWTLEAW** NCMSIAFPRD
TC75876	2.68e-14	NELRYLLKNK **RYFIVVDDIWNKSVW** ANILRALNKC
CD212839	3.08e-14	DIVREYLQGK **RYLLVIDDLWDPSAW** EIIKCAFPES
AW285775	4.06e-14	DTLREYLCDK **RYLIVLDDLWEVKHW** DIISCAFPKN
NP239124	4.64e-14	DVLMQGLEDK **RYFVVLDDLWKIDDW** NWIKTTAFPK
BZ342222	8.95e-14	EIIRKHLEGK **RFILVLDDVWEKDVW** INNIMEVFPT
BZ341506	8.95e-14	KNLQKLTNGK **RYLIVLDDVWNRDEA** KWEKLLTCLK
BZ369917	1.02e-13	ALMKMVEQKK **KFLLVMDDVWGEKVW** NDLLRVPLSY
Sb_RGA50	5.11e-13	KRIQEGLTRK **KFLIVMDDIWTESQN** QWDKIMDHLK
TC89319	6.49e-13	KSIAQRLKSK **RFLIVLDDIWECSSN** DEWEKLLAPF
Sb_RGA181	8.23e-13	RRLAIAVENA **SFFLVLDDIWQHEVW** TNLLRAPLNT
Sb_RGA125	1.17e-12	EKLKQKLKTR **KCLIVLDDVWDQEVY** LQMSDAFQNL
BH246154	1.47e-12	EVLQGILSKK **RYLVLLDDVWDAAAW** YEIRSAFVDD
TC85900	1.84e-12	KLRDILQKSQ **KFLLVLDDVWFEKSD** SETEWFQLLD
RHOH113F05G1	1.84e-12	KLRDILQKSE **KFLLVLDDVWFEKSD** SETEWFQLLD
NP853482	1.84e-12	RLRDILQKSE **KFLLVLDDVWFEKSD** SETEWFQLLD
TC79065	2.56e-12	AR **GYLIVIDDLWSSDQW** GIIRCCFPDN
BH246056	2.56e-12	SSLFSCSWLR **RYFVIIDDIWKASDW** EEIKGAFPNN
Sb_RGA75	3.19e-12	TQIEQILTSK **RFLLVLDDMWDTVNN** DGWDRLLAPF
AW564339	3.19e-12	EEIKNVLGTR **KCLFVLDDVWNKEVY** HQMMEDIFNT
AW672400	3.56e-12	FGVHFRVCCP **RYFIIIDDIWSERDW** NLLKCALPEN
BZ628476	4.41e-12	SFKMLYQLFG **RYLIVIDGLWETTSW** DIVSSAFPDD
BZ345488	4.90e-12	NAISNRLSGK **KFLLVLDDAWHDDRD** DWKQFLVHIR
CD209645	7.47e-12	AKLVDKLSGK **RFLLVLDDLWVNDEN** HQDLEEILSP
Sb_RGA130	1.38e-11	DLIEKRLQSK **QFLLVLDDMWTYHED** EWKKLLAPFK
NP239122	1.38e-11	RKLATAVENR **SVFLVLDDIWKHEVW** TNLLRTPLNT

Kin-2 motif consensus sequence **RYLIVLDDVWDxDVW**

A-4

ALIGNMENT OF CONSERVED MOTIF RNBS-B IDENTIFIED BY MEME

NAME	P-VALUE	SITES
BZ626449	2.80e-18	DATAWSAIRC **ALPENKNGSRVIATTRIEAVA** AACCSNDYEY
CD212839	4.42e-17	DPSAWEIIKC **AFPESHCGSRVLTTTRIVSVA** VACCNYQWKF
BH246056	2.30e-16	KASDWEEIKG **AFPNNNRGSRILITTRSTRTA** WACCSDSYYG
TC75876	2.99e-16	NKSVWANILR **ALNKCGRGSRIIITTRILDVA** QQADSVYKLQ
Sb_RGA55	3.41e-16	DKDAWLFLNH **AFVRNNCGSKVLITTRRKDVS** CLAVDHYRIE
BZ369917	3.88e-16	EKVWNDLLRV **PLSYGAPGSRVLVTTRNDEVA** RGINAQHLHR
AAM94297	1.06e-15	DISVWKMIRC **ALPDNMGGYVIITTTRNFKVA** EEIGGAYSMK
NP239124	2.19e-15	DDWNWIKTTA **FPKSNKKGSRILVTTRDASLA** KLCASIAGSF
BZ330329	2.47e-15	STAVWDSIIR **SFPRINNTSRIIVTTREENVA** RHCSSRPENV
TC80927	3.51e-15	TKSPWEIIQC **ALPNNGHTSKVIMTTRINSVG** QFSSTSDEGF
BH246133	4.97e-15	DKDAWLFLNY **AFVRNNCGSKVLITTRRKDIS** SLAVDNYAIE
AAM94306	4.97e-15	DITAWKMIKC **ALPDNCYGNKIITTTRILNIA** KQAGGAYNLE
AAM94294	4.97e-15	DVSVWEVIKC **ALPENDIGFAVITTTRNVDVA** DRSWWCLQVE
AW285775	8.75e-15	EVKHWDIISC **AFPKNSQQSRLIVTTRIEGVA** QACCKDHGRI
BH246154	1.36e-14	DAAAWYEIRS **AFVDDGTRSRIIITTRSQDVA** NLAKSTRTIL
AAM94295	1.52e-14	DISVWRMIKC **ALPHSDAGYIIITTTRNSDVA** EKVGSPYNMK
BZ628476	1.70e-14	ETTSWDIVSS **AFPDDTHCSRILITTNIEEVA** LECCDYESDA
Sb_RGA50	7.39e-14	SQNQWDKIMD **HLKAGAPGSGILITTRSKHVA** KAVRSTYQFC
Sb_RGA130	8.18e-14	EDEWKKLLAP **FKKVQTKGNMVIVTTRIPKVA** QMVTTIGCPI
Sb_RGA75	1.22e-13	NDGWDRLLAP **FRKGQTKGNMILVTTRSPSVA** QIVKVKPTDS
BZ341506	1.22e-13	DEAKWEKLLT **CLKQGDKGSTVLATTRDKEVA** RIMAIGASES
BI074536	1.49e-13	TIEEWDQIKK **CFPNNKKGSRIIVSSTQVEVA** SLCAGQESQA
AW564339	1.65e-13	NKEVYHQMME **DIFNTLRASRIIITTRREDVA** SLASSGCHLQ
BG050233	2.43e-13	TRHWNSLTA **PLSCCAPGSAVAVTTRSNKVA** RMVSTKVYHL
TC79065	4.33e-13	SSDQWGIIRC **CFPDNSLGSSIITTTRNDALP** TNHHCGSSKF
TC87218	4.76e-13	NRSMWKKVLA **PLRSAAIGSKVLVTTRMKLVA** EVLNAAHVVS
TC85900	1.20e-12	SETEWFQLLD **PFVSKQMGSKVLVTSRRETLP** AAVFCDQQQV
TC90621	1.44e-12	DVNKWGKLKS **SVQHGGSGSAVLTTTRDRVVA** KLMADTTHEP
Sb_RGA125	2.94e-12	WDQEVYLQMS **DAFQNLQSSRIIITTRKNHVA** ALAHPTRRLD
BZ367728	3.21e-12	SVVEWGMIIQ **SLPKMENASRILITTRERNIA** KHCSRNEESI
BM325057	8.25e-12	**TRPKHNSESKIVLTTRIEDVC** DRMDVRRKLR
TC76961	1.25e-11	IDAWNRIKVT **FQDSGKDDSCVVVTTRNQTLA** KYCSPPSHIH
NP853482	1.48e-11	SETEWFQLLD **PLISKQSGSKVLVTSRRAMLP** AAICCEQEQV

RNBS-B motif consensus sequence ALPxNxxGSRILVTTRIxxVA

A-5

ALIGNMENT OF CONSERVED MOTIF RNBS-C IDENTIFIED BY MEME

NAME	P-VALUE	SITES
BZ626449	2.54e-18	AACCSNDYEY **VYKMKALGTEDSRRLFFKRIF** GSEDTCPSYL
AW285775	2.54e-18	AQACCKDHGR **IHYMKPLSDADSRKLFFRRIF** GTEDTCPPQF
CD212839	1.17e-17	VACCNYQWKF **VYRMKPLDDYHSRQLFLRRIF** GSGDRCPEPF
BH246133	7.31e-17	DISSLAVDNY **AIELKTLQYAESWELFCKKAF** RASRDNQCPE
BZ342222	1.25e-16	EVASLATGNC **AIKLEPLGEKHSWKLFCKAAF** RNSDDKWCPS
Sb_RGA55	6.44e-16	DVSCLAVDHY **RIELKTLQYAESWELFCKKAF** VALKDSQCPE
BZ331922	8.20e-16	QVASIMGTLA **PHELKCLGEDDSWTLFSNKAF** SNGLQEQSEF
AAM94295	9.24e-16	NSDVAEKVGS **PYNMKPLSQNNSRKLLYKRIF** GNEGKDNNED
BG557168	1.67e-15	GTLP **HHELACLSDGDSWELFSKKAF** SKGVQKQEEL
TC86205	2.10e-15	HEAHHDH **VYEITPLSTDNSKCLFFKRIF** GSEHICPPHL
NP853482	5.76e-15	PAAICCEQEQ **VIHLENMDDADFLALFKHHAF** SGAKIGDQIL
AAD27570	5.76e-15	KICSSPFHDL **VFKLRMLSEDDSKRLFFRRIF** GSEDKCPHQL
TC75876	7.16e-15	ILDVAQQADS **VYKLQALSAGDSRKLFFLRIF** GNENRCLPKE
AAM94306	1.10e-14	ILNIAKQAGG **AYNLEPLSMNNSRKLLYRRIF** GTDSKDNNED
RHOH113F05G1	2.80e-14	PAAVHCELEQ **VVHLENMDDADFLALFKHHVF** SGPKIGDLLY
BZ628476	5.11e-14	LECCDYESDA **IFKMETLGGNHSTELFFNRVF** GFKHECSKQL
BG050233	1.34e-13	NKVARMVSTK **VYHLKCLSDEDCWRVCQRRAL** PNSDANVDQE
TC80927	1.48e-13	QFSSTSDEGF **IYQMKPLSRNDSENLFLKRTL** CAEDKFPVQL
TC85900	1.62e-13	PAAVFCDQQQ **VVHLEKMDDANFLALFKHHAF** SGAKIGDQLL
AAM94297	4.86e-13	NFKVAEEIGG **AYSMKALCHESSRKLFYTRIF** GNEEKYKCPD
BZ329687	1.39e-12	QQLRQSQAVI **VYQLEPLSLTDSKKLFCQIFG** SEDKCPPDNL
BH246154	1.39e-12	DVANLAKSTR **TILLKPLPEKEAWCLFCNTTF** REDADRECPQ
NP239124	1.79e-12	ASIAGSFHSL **VYCLEPLQDHHAKELLLKKTN** RSHQALKIGE
TC83499	2.30e-12	RFKFPTLVKQ **TYEMQLLDEAAALSVFCRAAF** DQESVPQTAD
TC79065	2.50e-12	HHCGSSKFVH **NHKISLLSDNEAKELFLKKAF** SSRNDYPQHL
BZ349832	3.78e-12	HPGHF **VYKVASLKHLDSRTLFLRRTF** GSEDNFPHDL
BH246056	4.45e-12	ACCSDSYYGL **VHEMKPLSETDSERLLLAKAV** GSVDGCVPNN
Sb_RGA130	4.82e-12	VAQMVTTIGC **PIRLERLSDEECMRFFQECVF** GDQQTWEGHT
Sb_RGA80	7.20e-12	SVVKMIATMD **PVHLDGLEDDDFWLLFKSCVF** GDEKYEGHGN
AW564339	7.20e-12	DVASLASSGC **HLQLQPLGSSYALDLFCRRAF** NNTADRKCPQ
BZ330329	9.88e-12	ARHCSSRPEN **VYXLNVLQYKDALDLFTKKVM** IRYISSWIVL
Sb_RGA75	1.46e-11	QIVKVKPTDS **TIELEGLDQVAFREFFQSCVF** GDDNKSKDDH
BZ367728	3.63e-11	AKHCSRNEES **IYNLQVLNPWDSLDLFTRKVL** YISSLAPSFI

RNBS-C motif consensus sequence VYELKPLSDxDSRELFxKRAF

A-6

ALIGNMENT OF CONSERVED MOTIF GLPL IDENTIFIED BY MEME

NAME	P-VALUE	SITES	
AAM94295	1.17e-22	DAELTEVSER	**ILKKCAGVPLAIITMASLLAC** KPRNKMDWYE
AAM94306	1.17e-22	PDELVEVSEK	**ILKKCAGVPLAIITMASLLAC** KARNKMEWCK
AAM94294	1.17e-22	IEELAEVSDR	**ILKKCAGVPLAIITMASLLAC** KPRNKMDWYE
CD212839	1.60e-21	PEPFEVLCEK	**ILQKCGGLPLAIITIASLLAS** QQTRSIEQWE
AW285775	3.13e-21	PPQFTEVSSE	**ILKKCGGLPLAIVTMASSLAD** QPKEHWDYIQ
AAM94297	3.13e-21	DEHLTEVSHR	**ILNKCAGVPLAIITIASLLAN** KARDKMEWLE
BZ329687	1.50e-20	PDNLVEVAGK	**ILKKYGGVPLAIITMASMLAN** KTGKEINAHN
BZ330329	1.74e-20	HPELIHEAKM	**ILKKCNGLPLAIVTIGGFLAN** QPKTVLEWRK
BE596218	1.74e-20	YPTLIEEAKM	**ILKKCKGLPLAIVTIGGFLAK** QPKTPIVWRK
AAD27570	2.02e-20	PHQLKDVSVE	**IIKKCGGLPLAIITMASLLTT** KSDTRADWLK
TC86205	3.12e-20	PPHLEDISSE	**ILEKCSGSPLAIVTMASLLAN** KACTKQEWDR
TC75876	6.30e-20	PKELDKESKN	**ILRKCGGVPLAIITISSMLAS** KQETENTSEY
TC80927	6.52e-19	PVQLTGIKND	**IIEKCDGLPLAIVTLASMLAT** K
TC76961	1.19e-18	GDKTKGIVEK	**ILNKCGGLPLAILTIGAVLAN** KDTEEWENIY
BZ626449	1.91e-18	PSYLEEVSTG	**ILKRCGGLPLAIITLSSHLAT** QRDKLDRELW
AW564339	2.15e-18	PQELEDVAVS	**IVERCKGLPLAIISMGSLMSS** KKPTKHAWNQ
BG557168	8.27e-18	QEELITIGKL	**IVSKCKGLPLALKTMGGLMSS** KHQIKEWEAI
BH246133	1.27e-17	PENLRFFAEK	**IVDKCQGLPLAIVTIGSTLSY** HELEEERWAF
WS110C06B1	1.42e-17	DDEFTKVAET	**ISKKCSGVPLAIVTLAKMLAT** KMGGKKEWHK
BZ349832	6.67e-17	PHDLEELSTK	**ILKKCAGLPLVIVCISSILAT** KGKEATEWEK
TC81018	1.96e-16	SDKEIIHHGE	**LWRRCGGQPLAIVTMAGLVAC** NQNKPTKYWD
BM325057	2.38e-16	SPEIRQQAQA	**LAMKCGGLPLALITVGRAMAS** KRTAKEWKHA
BH246154	3.16e-16	PQHLEHWALR	**ILNKCSGLPLAIVSVGNVLAL** KEKSEFAWKS
BZ628476	3.81e-16	SKQLKECSEE	**IIRTCGGLPLAIISIASILAI** QPDNLELWRH
BZ331922	4.59e-16	QSEFSTVGRR	**IVNKCKGLPFAFKAMGGLMSS** KPRVQQWEGI
TC90621	9.57e-16	DAKLVEMVGD	**IAKRCAGSPLAATAVGSLLQT** KTSVDEWNAV
BG050233	1.15e-15	DQELVEIGEK	**IAKKCQGLPLAAEAAGSALST** STSWKHWDEV
BZ349019	2.33e-15	SDELDVVVDK	**IVHRCVGSPLAAKAFGSMLST** KSSIQEWKDM
Sb_RGA75	3.03e-15	HKELDDIGEE	**IMKKLKGSPLAAKTVGRLLRN** NLDQNHWKRV
TC79065	6.03e-15	PQHLEDVFAK	**VLRRCGGLPLAVVSIATKLAH** KQSRDEWEKH
Sb_RGA130	1.78e-14	HTNLHYYGCK	**IVKRLKGFPLAVKTVGRLLKA** ELTADHWRRV
TC77858	2.47e-14	HTQIPALARQ	**VAAECKCLPLALVTVGRAMSN** KRTPEEWSNA
CD211851	3.99e-14	YRKIDRVTKK	**VVNICGGLPLALVSMAGYVGC** NKKPEELLKH

GLPL motif consensus sequence ILKKCGGLPLAIVTIGSLLAS

A-7

ALIGNMENT OF CONSERVED MOTIF RNBS-D IDENTIFIED BY MEME

NAME	P-VALUE	SITES
WS110C06B1	4.86e-32	YYNLPPHLRA **CLLYMSVFPEDYEIRRDRLVWRWIAEGFV** QFEDSKVESL
AAM94306	1.40e-31	YFDLPYHLRT **CLLYLSVFPEDYKISKNRLIWMWIAEGFI** QSGRHWGTLF
AAM94294	1.40e-31	YYNMPSHLRT **CLLYLSMFPEDYEVEKDRLIWMWIAEGFI** HCEKQGKSQY
BZ337854	5.69e-31	YNDLPTNLKT **CLLYLSIFPEDYVIERERLVRRWIAEGFI** CEERGLSKQE
AAM94295	1.01e-30	YYNMPSHLRT **CLLYFSVFPEDYKIEKHRLIWMWIAEGFI** QCEKHGESLF
TC86205	1.47e-30	FDDLPHHLKT **CLLYLSIFPEDYEIERDQLVKRWIAEGFI** NMEGGQDLEE
BZ338669	1.77e-30	HHLPSRLKP **CFLYLSIFPEDYEIKRSHLVHRWIAEGFV** RAKVGTTIDE
AAM94297	6.25e-30	YYDLKYHLRV **CLLYLSMFPEDYPITKNHLIWMWIAEGFV** QCEQGKSLFE
AAD27570	3.42e-29	YNHLPHHLKT **CLLYLSMFPEDYVIKRDYLVRRWVAEGFI** SAHGRKNLED
TC76169	6.56e-29	YYDLPAHLKT **CLLYLSVFPEDYEIVKDRLIWRWIAEDFV** PPGEGGQSSF
TC79065	1.20e-27	YNDLQPQLKS **CLLYLSIFPENSEIETKRLVRRWIAEGFI** AGTGSKEETA
BZ346314	1.30e-26	YYDLTPQLKT **CLLYLSIFPEDYQINKLRLIERWIAKGFV** QQGDGRQSLH
BZ329687	4.30e-26	YYDLPSHLMN **CFLYLSLFPEDYMIQIRALIWKWIGEGFV** RKEQGKTLYE
CD211851	7.21e-26	YNDMPAEIKT **CSLYLSIFPKGSRISRKRLTRRWIAEGFV** SEKQGMSMED
BM323307	8.20e-26	YNDMPAEIIT **CSLYLGIFPKGSRISRKRLIRRWIAEGFV** SEKDGMSVED
BH246154	1.20e-25	IDDLPYHLKR **CFLYCSIYPEDFFVKRKILIRKWIAEGFV** EEKNHATMED
BE596218	1.75e-25	YDGLPYHLKS **CFLYMSIFPEDYSISRRRLVHRWKAEGYS** SEVRGKSKGE
TC79359	4.72e-25	YIHLADELKQ **CFTFCSIFPKGYGIQKDRLIAQWIAHGFI** NAMNGEQLED
BZ349832	7.66e-25	YDDLPQHLKV **CLLYLSAFREDYAIRRDRLTRRWITEGFV** DEKPGMSMQE
TC75876	1.10e-24	YHDLPLHLRT **CLLYLSLYPEDYKIMTHDLVWKWIGKGFV** VIKQGMNMFE
NP853482	1.57e-24	YKKLDPRLQR **CFMYCSLFPKGHRYKPDELVHLWVAEGFV** GSCISGRRTL
BZ626449	2.23e-24	YTNLPHCLKA **CVLYLGMYPEDHEISKNDLVRQWVAQGFI** SKAGGQDAED
TC90621	5.02e-24	YNGLPPHIRQ **CFAFCAIFPKDYEIDVEKLIQLWMANGFI** PEQHGVCPEI
TC80849	7.07e-24	YVDLPSHLKE **CFLHCSLYPEEYPIQRFDLVRRWIAEGIV** NPRDNELLEE
TC85900	9.92e-24	YKKLDPRLQR **CFLYCSLFPKGHKYKPDELVHLWVAEGLV** GSCNLSSMTI
BZ628476	2.17e-23	YNSLPCHLKT **CLLYLSMYPEGYTFFKADLVKQWSAEGFI** IPGEEKNCDE
BG556059	2.17e-23	FRTCPDFLKP **CIFYLSIFPRGHRIRRRRLVRRWIAEGYA** RDTDKISADE
BM317647	2.70e-23	FVSCPDSLKP **CIFYLSIFPVNHKIRRRRLVRRWIAEGYS** TDTKE
BE355823	3.36e-23	INYLPGNVKN **CFLYCGLFPEDHQIRGEEIIRLWITEDFI** EERGPTSITM
TC81018	9.89e-23	YNDLHGDLKT **CLLYLAMFPKGCKTSRKCVTRRWIAEGFV** TKKYGLTEEE
BG557168	1.10e-22	YMHLSSEMKQ **CFAFCAVFPKDYEMDKDKLIQLWMANNFI** HADGTTDFVQ
BZ342222	2.30e-22	LEDLPYELKN **CFLYCAIFPEDQELTRRTLMRHWITSGFI** KEKDNRTLEQ
BH245455	2.55e-22	YFDLPHHLKS **CLLYLSVFPEDFSIDCRELILLWVAEGLI** PGQDRESMEQ

RNBS-D motif consensus sequence CFLYLSIFPEDYEIxRDRLIRRWIAEGFI

A-8

ALIGNMENT OF CONSERVED MOTIF MHDV IDENTIFIED BY MEME

NAME	P-VALUE	SITES			
AAM94306	5.08e-28	RSMIQPIHDT	**DTGLIKQCRVHDMILDLICS**	**LSSEENFV**	TILTDVDGTS
AAD27570	2.10e-27	RSLIQPVDFQ	**YDGRVYTCRVHDVILDLITC**	**KAVEENFV**	TVVTNGKQML
AAM94295	4.63e-26	RSMIQPIHGY	**NNDTIYECRVHDMVLDLICS**	**LSSEGNFV**	TILNGTDHIP
AAM94294	9.50e-26	NRSMIQPIYG	**VSSNVYECRVHDMVLDLICS**	**LSSEANFV**	TILNGMDQMS
BM323307	3.80e-25	RKMIRPVEHS	**SSGRIKQCVVHDMVLEHIVS**	**KASEENFI**	TVVGGHWLKN
BH245455	5.32e-25	SLVQPTKVGV	**DGTNVKQCRVHDVILEFIVS**	**KAVEDNFV**	TIWNGDGFSR
BZ338669	5.95e-25	RSMIQSSELG	**MEGSVKTCRVHDIMRDIIVS**	**ISREENFV**	HLVQSNGNNV
TC76169	1.15e-24	RSLIQPADMD	**DEGTPISCRVHDMVLDLICN**	**ISREESFV**	ATVLDDARQN
BZ337854	1.98e-24	KSMVQPVDVG	**YDGKARACQVHDMMLELIIS**	**KSIEDNFI**	SLVGHGQTDL
BZ626449	3.36e-24	RSIIQPAHTD	**SNNDVLSCRVHDMMLDLIIH**	**KCREENFA**	TASDDIEGLE
TC81018	3.87e-23	RKLIRPVDHS	**SNGKLKTFQVHDMVLDYIAS**	**KAREENFI**	TVIGGHWMMP
AAM94297	5.72e-23	TSMIQPVYDR	**HEAMIEHCRVHDMVLEVIRS**	**LSNEDNFV**	TILNNEHSTS
BZ346314	4.60e-22	SLIQPADLDE	**DEMNLFSCRVHDMVLDLICS**	**LSRDESFA**	TTLNGDCKEI
TC79065	5.53e-20	RNLVQPLDLN	**HDNIPRRCTVHPVIYDFIVC**	**KSMEDNFA**	TLTDAQHVPN
BZ342222	6.71e-17	RSLLQVVIKN	**ASGRVKRCRMHDVIRHLAIE**	**KAAKECFG**	IIYEGYGNFS
BG556059	5.76e-16	QKTYSVTTTF	**GGRRMTLCQVNSFVREYIIS**	**RQMEENLV**	FELGGSCTLT

MHDV motif consensus sequence **DEGRVKxCRVHDMVLDLICSKSREENFV**

A-9

ALIGNMENT OF CONSERVED MOTIF RNBS-IV IDENTIFIED BY MEME

NAME	P-VALUE	SITES
AAD27570	1.84e-17	KNCDVEEMNM **ILSLSYNHLPHHLKT** CLLYLSMFPE
TC76169	1.83e-16	SNPDMENMRK **ILSLSYYDLPAHLKT** CLLYLSVFPE
Sb_RGA130	3.25e-15	YQANDDDIMP **ALKLSYNYLPFHLQQ** CFAYCALFPE
AAM94306	2.56e-14	NNSALENMRK **ILAFSYFDLPYHLRT** CLLYLSVFPE
TC86205	2.88e-14	KDPDVEEMRR **ILSLSFDDLPHHLKT** CLLYLSIFPE
AAM94295	3.65e-14	NSIDVENMRK **ILSFSYYNMPSHLRT** CLLYFSVFPE
BZ329687	3.65e-14	GSTNVKNMRR **ILSVSYYDLPSHLMN** CFLYLSLFPE
AAM94294	3.65e-14	NNLDVENMRK **ILSFSYYNMPSHLRT** CLLYLSMFPE
BH245455	5.80e-14	KDSPIDKMKR **ILLLSYFDLPHHLKS** CLLYLSVFPE
TC90621	1.27e-13	ICDDETEILP **ILKLSYNGLPPHIRQ** CFAFCAIFPK
TC75876	1.41e-13	TSSDVIDMRR **ILSVSYHDLPLHLRT** CLLYLSLYPE
BH246154	1.57e-13	TDHGIGQVSS **ILNLSIDDLPYHLKR** CFLYCSIYPE
BZ337854	2.68e-13	KNRSLEGMNS **ILCLSYNDLPTNLKT** CLLYLSIFPE
BZ349019	3.30e-13	ICDERTEIFP **ILKLSYDDLPSDMKQ** CFAFCAVFPK
BZ628476	4.50e-13	NLTSEVKLRE **IVSLSYNSLPCHLKT** CLLYLSMYPE
Sb_RGA80	4.98e-13	LQQGPDDIIP **ALKVSYIHLPFHLQR** CFSYCAFFPE
Sb_RGA75	6.75e-13	LQTGDSDIMP **ALKLSYDFLPFHLQH** CFSYCALFPE
TC89319	8.24e-13	EENHDNDIIP **ALKISYDYLPFHLKK** CFSCFCLFPD
BE596218	2.39e-12	MNPELGIIRA **ILMKSYDGLPYHLKS** CFLYMSIFPE
TC79065	4.19e-12	RPEGLDGLKQ **ILNLSYNDLQPQLKS** CLLYLSIFPE
AAM94297	7.23e-12	DSTDVENMRK **ILAYSYYDLKYHLRV** CLLYLSMFPE
BZ423689	1.03e-11	**ALKLSYDYLPDSLQQ** CFRYCCLFPK
BZ626449	1.74e-11	LNPTLEGMRQ **ILSMSYTNLPHCLKA** CVLYLGMYPE
TC76961	2.07e-11	NNPSLDALRR **VVSLSYNHLPSRLKP** CFLHLSIFPE
BZ349832	2.46e-11	SNDGLSWLWQ **AFEVSYDDLPQHLKV** CLLYLSAFRE
TC80849	3.75e-11	VSPVLPEVPQ **AVYVSYVDLPSHLKE** CFLHCSLYPE
WS110C06B1	6.67e-11	NTLDVKNMRM **VTSLGYYNLPPHLRA** CLLYMSVFPE
BZ342222	7.85e-11	TNNVIRGVDI **ILKVSLEDLPYELKN** CFLYCAIFPE
BG557168	7.85e-11	DRVGKDEVLS **ILKLSYMHLSSEMKQ** CFAFCAVFPK
Sb_RGA125	8.51e-11	ELSNNDHVRA **VLNLSYNDLSGDLRN** CFLYCALFPE
BZ346314	8.51e-11	KKQSWYGYEE **DLLLSYYDLTPQLKT** CLLYLSIFPE
BH246133	2.77e-10	NNPELNWISN **VLNMSLNDLPSYLRS** CFLYC
TC81018	2.99e-10	NSLTLEGVKR **ILDCCYNDLHGDLKT** CLLYLAMFPK
CD211851	5.08e-10	EGLNQEEAGR **IISYCYNDMPAEIKT** CSLYLSIFPK
BZ340437	8.52e-10	FPGKYRNCYT **ALRLSCHHSPVHLRT** CFRYCSIFPP
BE355823	1.14e-09	NNPDLNAVRN **ALDLSINYLPGNVKN** CFLYCGLFPE
TC79359	1.63e-09	VQSIKDRVFA **SLKLSYIHLADELKQ** CFTFCSIFPK
NP853482	3.29e-09	KLRDLSEPLT **ILLWSYKKLDPRLQR** CFMYCSLFPK
TC85900	1.04e-08	KLRDLSEPFT **VLLWSYKKLDPRLQR** CFLYCSLFPK

RNBS-IV motif consensus sequence **ILSLSYNDLPSHLKT**

A-10

ALIGNMENT OF CONSERVED MOTIF RNBS-VI IDENTIFIED BY MEME

NAME	P-VALUE	SITES
AAM94297	3.08e-26	MWIAEGFVQC **EQGKSLFELGECYFNELINTSMIQPVY** DRHEAMIEHC
AAM94295	4.31e-26	WIAEGFIQCE **KHGESLFDLGESYFNELISRSMIQPIH** GYNNDTIYEC
AAM94294	5.99e-26	WIAEGFIHCE **KQGKSQYELGENYFNELINRSMIQPIY** GVSSNVYECR
TC86205	9.77e-26	RWIAEGFINM **EGGQDLEEIGENYFNDLINRSMIQPMK** IKCDGR
AAD27570	1.30e-23	RWVAEGFISA **HGRKNLEDEGECYFNELINRSLIQPVD** FQYDGRVYTC
AAM94306	5.70e-23	WIAEGFIQSG **RHWGTLFACGESYFNELINRSMIQPIH** DTDTGLIKQC
TC76169	9.59e-23	WIAEDFVPPG **EGGQSSFELGLSYFNDLVNRSLIQPAD** MDDEGTPISC
BZ346314	9.59e-23	WIAKGFVQQG **DGRQSLHEIGQSYFNELLNRSLIQPAD** LDEDEMNLFS
BH245455	1.67e-20	LWVAEGLIPG **QDRESMEQLGRSYLNELINRSLVQPTK** VGVDGTNVKQ
BZ626449	5.63e-20	QWVAQGFISK **AGGQDAEDIAVEYFNEIVNRSIIQPAH** TDSNNDVLSC
BZ337854	1.07e-19	RWIAEGFICE **ERGLSKQEVAENNFYELINKSMVQPVD** VGYDGKARAC
BZ349832	2.25e-19	RWITEGFVDE **KPGMSMQEVADNNFTELIGRNMIQAVD** VDCFGEIHAC
BZ342222	1.42e-18	HWITSGFIKE **KDNRTLEQVAEEYLNDLVNRSLLQVVI** KNASGRVKRC
TC75876	2.85e-18	KWIGKGFVVI **KQGMNMFEAGEDYVHELINRSLILPTF** DNKSKKAKF
NP853482	3.47e-18	VAEGFVGSCI **SGRRTLEDVGMDYFNDMVSGSLFQMVS** QRYFVPYYIM
BM323307	4.22e-18	RWIAEGFVSE **KDGMSVEDVAETYFGHLVRRKMIRPVE** HSSSGRIKQC
BZ338669	6.85e-18	RWIAEGFVRA **KVGTTIDEVGKEYFDELISRSMIQSSE** LGMEGSVKTC
TC81018	1.47e-17	RWIAEGFVTK **KYGLTEEELAETYFNQLLRRKLIRPVD** HSSNGKLKTF
TC79065	1.95e-17	VRRWIAEGFI **AGTGSKEETAISYLNELIGRNLVQPLD** LNHDNIPRRC
WS110C06B1	3.55e-16	IAEGFVQFED **SKVESLFELGESYVDEFVNRSMIQLLK** KKKKKLETSS
TC85900	3.55e-16	VAEGLVGSCN **LSSMTIEDVGRDYFNEMLSGSFFQLVS** ETEYYSYYIM
BE596218	7.11e-16	RWKAEGYSSE **VRGKSKGEIADAYFMELIERSMVLPSK** ESIGSRKGIS
BG050233	1.00e-15	LWTAQGFVDA **EGDCSLEAIANGYFNDLVSKCFFHPSP** SHAISEGKLV
BE355823	3.00e-15	ITEDFIEERG **PTSITMEEVGAEYLNEIAQRSLLQVVQ** RDAYGRSEIF
TC90621	3.54e-15	LWMANGFIPE **QHGVCPEITGKKIFMDLVSRSFFQDVN** KVPFEVYDIE
BM324406	4.92e-15	MALGFIQPPT **DEGKGMEDLGQKYFDDLLSRSFFGTAN** KDQQTYYFLD
BG557168	5.34e-15	LWMANNFIHA **DGTTDFVQKGEFIFSELVWRSFIQDVD** VKIFDEYHFA
TC80849	6.83e-15	RWIAEGIVNP **RDNELLEESAEEYYVELISRNLLQPDP** ESVERCWITH
TC89319	1.82e-11	WHSIGIIDYS **RQNKKMEEIGSDYLDELVDSGFLIKGD** DNYYVMHDLL

RNBS-VI motif consensus sequence KGGKSLEELGESYFNELINRSLIQPVD

A-11

ALIGNMENT OF CONSERVED MOTIF PRE-P-LOOP IDENTIFIED BY MEME

NAME	P-VALUE	SITES
AAM94294	3.69e-33	VNNGVDKPTT **TTVVDPRLFAQFKEAKELVGIDETRDELIKVLMDGNGVPFQ** QGKVVSIVGF
AAM94297	1.11e-28	YKIDGVGGAR **PDVVDPRLLAHYTAVTELVGIDDARDELIKVLTDDGSQEAS** KQHGRVVSIV
TC81885	1.72e-28	RYKVHAITPT **KTSVDPRIAALYTKASSLVGIDEPKEELISMLTKEDGGRSS** AEQWIVSIVG
OX158E07B1	3.64e-28	YKLDEKIAAA **PTIIDPRLIATYKEVSQLIGVDKSRDDLISMLNLLQPDDDA** PDKKITKKVS
NP853482	2.14e-27	NTTALGCPAV **PTTIVPLTTVTSLSTSKVFGRDKDRDRIVDFLLGKTAADEA** SSTRYSSLAI
TC85900	3.94e-27	NTTGLGWPNV **PATIVPPTTVTSLSTSKVFGRDKDRDRIVDFLLGKTAADEA** SSTRYSSLAI
AAM94295	4.36e-27	DVSLGVDKPS **TAAVDPRLFSQYTEIEELVGIVETRDELINIVMEENEVPIQ** KGKIVTIVGF
BZ334356	4.05e-23	VGNIIAAKPD **IVPVDPRLEAMYRRATELVGIGGPKNELAKRLLEEDCSSSS** RQQSNIISIV
RHOH113F05G1	2.53e-22	**TSNHCSSNHSDILSTSKVFGRDKDRDHIVDFLLGKTAADEA** SSTRYSSLAI
BZ343608	4.12e-22	TPSISSDVTL **DMELTRNLTALYVEETQLFGLDKQKEKLMDLIANPKVPVDM** EPGISKSGPR
AAD27570	4.12e-22	DDTVNFGGTN **VIPVDRRLPALYAELGGLVGISVPRDEVIKLVDDGAQGVKV** VSIVGCGGLG
TC76961	5.84e-20	TPSTSTNVIG **DTEFTRNFAALNVEEAQLVGLDEPKKKLMELIGILDEPKEH** ESSNTGPRVV

Pre-P-loop consensus sequence PTxVDPRLTALYLEASELVGIDKPRDELIDFLLDEDAADEA

APPENDIX B

B-1

SORGHUM NBS SEQUENCE ALIGNMENT OF AMINO ACIDS COVERING P-LOOP and KIN-2 MOTIFS FOR PHYLOGENETIC ANALYSIS

```
48          111
Apaf-1      GCGKSVLAAE AVRDHSLLEG ----CFPGGV HWVSVGKQD- ----KSGLLM KLQNLCTRLD
XA1         GIGKTTLAQL VCKD-LVIK- ----SQFNVK IWVYVSD--- ----KFDVVK ITRQILDHVS
TC89319     GIGKTTFTQH LVNHTRIKQ- ----CFHDIN IWICVST--- ----NFDVLK LTKEMLSCLP
TC89312     GIGKTTLLHA FNNDLEKKD- ----HNYQVV IFIEVSNSE- ----TLNTVE MQQTISDRLN
TC85900     GMGKSTLVQY VYND-KRIE- ----EGFDIR MWVCISR--- ----KLDVRR HTREIIESAT
TC81885     GLGKTTLAKA VYDKIKP--- -----QFDCT AFISVFR--- ----DPDIIK IFKDMLYELD
TC76961     GIGKTTLTKK VFDS-NDLS- ----DKFGTR AWITVSQ--- ----SFEQKE IFKEMVKHLF
Sb_RGA80    GMGKTTLAQL VYND-AVVQ- ----DHFNKR IWISVSI--- ----HFDEVR LTREMLDCLS
Sb_RGA75    GMGKTTLAQH IYN--KEAE- ----TYFDVR IWACVST--- ----DFSVPR LLKDILESKS
Sb_RGA55    GLGKTTIASS VYKN-QNIT- ----RTFNCH AWVTVSQ--- ----TYQVEE LLREIINQLI
Sb_RGA50    GWGKTTLAKL IFNDGETIE- ----KQFEVR LWVHVSQEFD -------FEK LIKKLFEAFA
Sb_RGA182   GVGKTTMAKL LFKDIQFNK- -------YSR VWVYVSE--- ----IFNLKK IGNSIISQVS
Sb_RGA130   GLGKTTFTQH IN---EDMK- ----SHFHVR VWVCISQ--- ----NFSASR LAQEIAKQIP
RPS5        GVGKTTLLTK INNKFSKID- ----DRFDVV IWVVVSR--- ----SSTVRK IQRDIAEKVG
RPS2        GVGKTTLMQS INNELITKG- ----HQYDVL IWVQMSR--- ----EFGECT IQQAVGARLG
RPP8        GIGKTTLARQ VFHH-DLVR- ----RHFDGF AWVCVSQ--- ----QFTQKH VWQRILQELQ
RPP5        GIGKSTIGRA LFSQLSSQF- ----HHRAFL TYKSTSGS-- --DVSGMKLS WQKELLSEIL
RPP1        GIGKTTIARF LFNQVSDRF- ----QLSAIM VNIKGCYPRP CFDEYSAQLQ LQNQMLSQMI
RPM1        GSGKTTLSAN IFKS-QSVR- ----RHFECY AWVTISK--- ----SYEIED VFRTMIKEFY
RP1D        GMGKSTLAQY VYND-KRIE- ----ECFDIR MWVCISR--- ----KLDVHR HTREIIESAK
RHOH113F    GMGKSTLAQY VYND-ERIK- ----EGFDVR IWVCISR--- ----KLDVRR HTRKIIESAT
PRF         GLGKTTLAKK IYNDPEVTS- -----RFDVH AQCVVTQ--- ----LYSWRE LLLTILNDVL
PIB         GLGKTTLVSG VYQS-PRLS- ----DKFDKY VFVTIMR--- ----PFILVE LLRSLAEQLH
NP853482    GMGKSTLAQY VYND-KRIE- ----EGFDIR MWVCISR--- ----KLDVRR HTREIIESAT
NP239124    GVGKTTLVRK VFHS-IDIV- ----GNFSSR AWITVSQ--- ----SFDKKE LLKELIKQLF
NP239123    GVGKTTLVRD VYETLEIKN- -----HFVEQ ALATFPP--- ----YSSASD ILKLILRDLK
NP239121    GLGKTTLVTN VYER-EKIN- ------FSAT AWMVVSQ--- ----TYTIEA LLRKLLMKVG
N           GVGKTTIARA IFDTLLGRMD --SSYQFDGA CFLKDIKEN- ----KRGMHS LQNALLSELL
MLA6        GLGKTTLARA VYEKIKG--- -----DFDCR AFVPVGQ--- ----NPDMKK VLRDILIDLG
M           GIGKTTTAKA VYNKISSHF- ----DRCCFV DNVRAMQEQ- --KDG--IFI LQKKLVSEIL
L6          GIGKTTTAKA VYNKISSCF- ----DCCCFI DNIRETQE-- --KDG--VVV LQKKLVSEIL
I2C-1       GMGKTTLAKA VYND-ERVQ- ----KHFGLT AWFCVSE--- ----AYDAFR ITKGLLQEIG
GPA2        GIGKTTLAAK LYSDPYIMS- -----RFDIR AKATVSQ--- ----EYCVRN VLLGLLSLTS
CD212839    GLGKTTLANE VYRDLRDNLS GNPEKSFSCK AIISVSQ--- ----RPDMVN LLKSLFTKVS
CD209645    GSGKTTLAQY VCDHENENRE ----KYFNPI MLIHVSE--- ----TFRVSD ILHDMLEAIT
BZ625990    GWGKTTLAKL IFNDGETIE- ----KQFEVR LWVHVSQEFD -------FEK LIKKLFEAFA
BZ369917    GIGKTTLARM VFND-NKIE- ----ENFEDR IWLSVNQ--- ----EVNEIS VLQSVLASFG
BZ367728    GLGKTTLVKE VYQS-QELS- ----DLFEKR ACVTIIR--- ----PFVLDE VLKSLAMQLR
BZ350423    GSGKTTLAKQ ICHDVKIKQ- ----HFKSTI FWVHVSEEFD -------VKE LIGKLFETIL
BZ345488    GSGKTTLAKL ICLDKRTKE- ----HFKDSI LWVHVSQEFD -------LEK LIGKLFESIA
BZ343608    GLGKTTLTKK VYDS-KDLG- ----DIFEIR AWIAVSQ--- ----SFDPKE LLKEMIKQLF
BZ341506    GLGKTTFIQL VYND-PAIQ- ----KHFQLQ RWCSVSD--- ----SFDIAN IASRICQTN-
BH246154    GVGKTTLVTN VFRE-VAAS- ----FHFDCA AWVSVSK--- ----NFTRED LLKRVLKELQ
BH246040    GLGKTTIASS VYKN-QNIT- ----RTFNCH AWVTVSQ--- ----TYQVEE LLREIINQLI
AAM94297    GLGKTTLANV VYQKIRT--- -----QFDCW AFVSVSQ--- ----TPDMRR LFEGILSELG
AAM94295    GLGKTTLAHA VFDKIRP--- -----GFDCC ASVSVSQ--- ----TPDLKK LLKGILYQLD
AAM94294    GLGKTTLAKV VYEKIRS--- -----LFHCC AFISVSQ--- ----TPDLKK LFKELLYDLD
AAD27570    GLGKTTIANQ VYINIAE--- -----KFDCQ AFVSLTQ--- ----NPDMVI IFQSILTQVK
```

SORGHUM NBS SEQUENCE ALIGNMENT OF AMINO ACIDS COVERING P-LOOP and KIN-2 MOTIFS FOR PHYLOGENETIC ANALYSIS (Continued)

```
QDES------ ------FSQR LPLNIEEAKD RLRILMLRKH PRSLLILDDV W
N--------- ------QSHE GISNLDTLQQ DLEEQMK--S KKFLIVLDDV W
ATENEENN-- ---------E TTTNLDQLQK SIAQRLK--S KRFLIVLDDI W
LPWNELE--- ---------- ---TVEKRAR FLAKALA--R KRFLLLLDDV R
NG-------- -------ECP CIDNLDTLQC KLRDILQK-S QKFLLVLDDV W
NKEYW----- ------DIHN IALGQHYLTD LVHEFLK--N KRYVVSCLVV P
GAESLHKLLE DHQG------ QQVLEVHLAD YLSKRLK--E TRYLIVLDDV W
DGV------- ------SKHD EIINLNKLQE ILEQSAK--S KRLLLVLDDM W
LHELSKG--- ---------- ---LTGTPET QIEQILT--S KRFLLVLDDM W
DQRAS----- -MAS----GF MTMNHMRLVE VIQSYLL--D KKYLIVLDDV W
DKDP------ ---------- GQPSLPYMSK RIQEGLT--R KKFLIVMDDI W
KTES------ ---------- -QITMQMIHT HLAELLA--G KNILIVLDDI W
KLDNEK---- ---------- ---ENESAED LIEKRLQ--S KQFLLVLDDM W
LGGMEWS--- ---------- -EKNDNQIAV DIHNVLR--R RKFVLLLDDI W
LSWDEKE--- ---------- ---TGENRAL KIYRALR--Q KRFLLLLDDV W
PHDG------ --------DI LQMDEYALQR KLFQLLE--A GKYLVVLDDV W
GQKD------ ---------- -IKIEHFG-- VVEQRLN--H KKVLILLDDV D
NHKD------ ---------- -IMISHLG-- VAQERLR--D KKVFLVLDEV D
KEAETQ---- -IPA----EL YSLGYRELVE KLVEYLQ--S KRYIVVLDDV W
KG-------- -------ECP RVDNLDTLQC KLRDILQE-S QKFLLVLDDV W
NG-------- -------ECP CIGNLDTLQC KLRDILQK-S EKFLLVLDDV W
EPSDR----- ---------- NEKEDGEIAD ELRRFLL--T KRFLILIDDV W
KGSSKKEELL ENRVSSKKSL ASMEDTELTG QLKRLLE--K KSCLIVLDDF S
NG-------- -------ECP CIDNLDTLQC RLRDILQK-S EKFLLVLDDV W
GDGSSKEHSR GLENNKVSGL QSKKVDGLMD VLMQGLE--D KRYFVVLDDL W
EED------- ---------F TLSKMEVTKE LLDKKLK--G KQYLVVIDGE V
GEQQ------ -----VPPNI DKLDVYDLKE KLKQKLK--T RKCLIVLDDV W
REKAN----- ---------- -YNNEEDGKH QMASRLR--S KKVLIVLDDI D
NP-------- ------HSDL AMLDANQLIK KLHEFLE--N KRYLVIIDDI W
RMDS--V--- ---------- GFTNDSGGRK MIKERVS--K SKILVVLDDV D
RIDSGSV--- ---------- GFNNDSGGRK TIKERVS--R FKILVVLDDV D
STDLKADDNL NQLQ--VKLK ADDNLNQLQV KLKEKLN--G KRFLVVLDDV W
D--------- ---------- --EPDYQLAD QLQKHLK--G RRYLVVIDDI W
GQ-------- -------TAD HTYDLPGLID IVREYLQ--G KRYLLVIDDL W
EDRHS----- ---------- DISGCKGLQA KLVDKLS--G KRFLLVLDDL W
DKDP------ ---------- GQPSLPYMSK RIQEGLT--R KKFLIVMDDI W
ANQNHEG--- ---------- FAGNKDLLER ALMKMVEQ-K KKFLLVMDDV W
GESFNRKDNI DFGIGIRK-- -LTETKLLTE ELGHLTK--R KRCLIVLDDL F
EQKS------ ---------- DLHAQQHMVD AISSKLR--G KKFLLVLDDA W
KKKA------ ---------- DRHTQQYMVN AISNRLS--G KKFLLVLDDA W
GAHSLKEFLE EHQG------ QVLEVKHLTN YLRGRLL--E RKYLVVLDDV W
---------- ---------E NDQGSENALK NLQKLTN--G KRYLIVLDDV W
RDVSAG---- -VPK----DV EETSYRSLVE VLQGILS--K KRYLVLLDDV W
DQRAS----- -MAS----GF MTMNHMRLVE VIQSYLL--D KKYLIVLDDV W
KD-------- -------INE ETRDVRHFID AIGKFLQ--T KRYCIVIDDI W
KKYE------ ------DINE KPLDEGQLVN ELRKFLR--R KRYFIVIDDI W
KN-------- -------INA ETLDERRLIN VLREFLI--P KRYLVVIDDI W
KD-------- ---------- ---------E CDSTSSC--D KE-------- -
```

B-2

SORGHUM NBS SEQUENCE ALIGNMENT OF AMINO ACIDS COVERING KIN-2 AND GLPL MOTIFS FOR PHYLOGENETIC ANALYSIS

```
47          130
Apaf-1      SLLILDDVWD S--------- --WVLKAFDS Q--------- -----CQILL TTRDKSVTDS
XA1         FLIVLDDVWE I-------RT DDWKKLLAPL RPNDQVNSSQ EEATGNMIIL TTRIQSIAKS
TC89319     FLIVLDDIWE CS------SN DEWEKLLAPF K--------K DETSGNVILV TTRFPKIVEM
TC87218     FLLVLDDVWI DEGKTEKENR SMWKKVLAPL R--------- SAAIGSKVLV TTRMKLVAEV
TC85900     FLLVLDDVWF EKS----DSE TEWFQLLDPF VS-------- -KQMGSKVLV TSRRETLPAA
TC80927     YLVVIDDVWT K--------- SPWEIIQCAL P--------- NNGHTSKVIM TTRINSVGQF
TC79065     YLIVIDDLWS S--------- DQWGIIRCCF P--------- DNSLGSSIIT TTRNDALPTN
TC76961     YLIVLDDVWT I--------- DAWNRIKVTF Q--------D SGKDDSCVVV TTRNQTLAKY
TC75876     YFIVVDDIWN K--------- SVWANILRAL N--------- KCGRGSRIII TTRILDVAQQ
Sb_RGA80    LLLVLDDMWG RQ------DK SRWEKLLAPL R--------C SLLKGSVILV TTRNHSVVKM
Sb_RGA75    FLLVLDDMWD TV------NN DGWDRLLAPF R--------K GQTKGNMILV TTRSPSVAQI
Sb_RGA55    YLIVLDDVWD K--------- DAWLFLNHAF VR-------N NCG--SKVLI TTRRKDVSCL
Sb_RGA50    FLIVMDDIWT ES-------Q NQWDKIMDHL K--------- AGAPGSGILI TTRSKHVAKA
Sb_RGA130   FLLVLDDMWT Y-------HE DEWKKLLAPF K--------K VQTKGNMVIV TTRIPKVAQM
RPS5        FVLLLDDIWE K--------- VNLKAVGVPY PS-------K DNG--CKVAF TTRSRDVCGR
RPS2        FLLLLDDVWE E--------- IDLEKTGVPR PD-------R ENK--CKVMF TTRSIALCNN
RPP8        YLVVLDDVWK K--------- EDWDVIKAVF P--------- -RKRGWKMLL TSRNEGVGIH
RPP5        VLILLDDVDN ---------- LEFLKTLVGK AE-------- WFGSGSRIIV ITQDRQLLKA
RPP1        VFLVLDEVDQ ---------- LGQLDALAKD TR-------- WFGPGSRIII TTEDQGILKA
RPM1        YIVVLDDVWT T--------- GLWREISIAL P--------- DGIYGSRVMM TTRDMNVASF
RP1D        FLLVLDDVWF EKS----HNE TEWELFLAPL VS-------- -KQSGSKVLV TSRSKTLPAA
PRF         FLILIDDVWD Y--------- KVWDNLCMCF SD-------V SNR--SRIIL TTRLNDVAEY
PIB         CLIVLDDFSD T--------- SEWDQIKPTL FP-------- LLEKTSRIIV TTRKENIANH
NP853482    FLLVLDDVWF EKS----DSE TEWFQLLDPL IS-------- -KQSGSKVLV TSRRAMLPAA
NP239124    YFVVLDDLWK I--------- DDWNWIKTTA FP-------K SNKKGSRILV TTRDASLAKL
NP239123    YLVVIDGEVS S--------- TEWKNILGAL P--------- -NVAGSKVVR MSKENLEDPP
NP239122    VFLVLDDIWK H--------- EVWTNLLRTP LN-------- -TSSTTIIVL TTRNDIVARV
NP239121    CLIVLDDVWD Q--------- EVYLQMS-DA FQ-------N LQS--SRIII TTRKNHVAAL
N           VLIVLDDIDN K--------- DHYLEYLAGD LD-------- WFGNGSRIII TTRDKHLIEK
MLA6        YLVIIDDIWD E--------- KLWEGINFAF SN-------- RNNLGSRLIT TTRIVSVSNS
M           ILVVLDDVDE K--------- FKFEDILGCP KD-------- FDS-GTRFII TSRNQNVLSR
L6          ILVVLDDVDE K--------- FKFEDMLGSP KD-------- FIS-QSRFII TSRSMRVLGT
I2C-1       FLVVLDDVWN DN-------Y PEWDDLRNLF L--------- QGDIGSKIIV TTRKESVALM
GPA2        YLVVIDDIWT T--------- EAWDDIKLCF PD-------C DNG--SRILL TTRNVEVAEY
CD212839    YLLVIDDLWD P--------- SAWEIIKCAF P--------- ESHCGSRVLT TTRIVSVAVA
BZ628476    YLIVIDGLWE T--------- TSWDIVSSAF P--------- DDTHCSRILI TTNIEEVALE
BZ626449    FFIVIDDIWD A--------- TAWSAIRCAL P--------- ENKNGSRVIA TTRIEAVAAA
BZ342222    FILVLDDVWE K--------- DVWINNIMEV FP-------T NCT--SRFVF TSRKFEVASL
BH246154    YLVLLDDVWD A--------- AAWYEIRSAF VD-------D GTR--SRIII TTRSQDVANL
BH246056    YFVIIDDIWK A--------- SDWEEIKGAF P--------- NNNRGSRILI TTRSTRTAWA
AW564339    CLFVLDDVWN K--------- EVYHQMMEDI FN-------T LRA--SRIII TTRREDVASL
AW285775    YLIVLDDLWE V--------- KHWDIISCAF P--------- KNSQQSRLIV TTRIEGVAQA
AAM94306    YLIVIDDIWD I--------- TAWKMIKCAL P--------- DNCYGNKIIT TTRILNIAKQ
AAM94297    YCIVIDDIWD I--------- SVWKMIRCAL P--------- DNMGGYVIIT TTRNFKVAEE
AAM94295    YFIVIDDIWD I--------- SVWRMIKCAL P--------- HSDAGYIIIT TTRNSDVAEK
AAM94294    YLVVIDDIWD V--------- SVWEVIKCAL P--------- ENDIGFAVIT TTRNVDVADR
AAD27570    ---------- ---------- ---------- ---------- -----SRIIV TTRIGTVAKI
```

SORGHUM NBS SEQUENCE ALIGNMENT OF AMINO ACIDS COVERING KIN-2 AND GLPL MOTIFS FOR PHYLOGENETIC ANALYSIS (Continued)

```
VMGPK----- YVVPVESSLG KEKGLEILSL FVN------- ------MKKA DLPEQAHSII KECKGSPLVV
LGTVQSI--- ---KLEALKD DDIWSLFKVH AFG------- --NDKHDSSP GLQVLGKQIA SELKGNPLAA
VKKETNP--- --IDLRGLDP DEFWKFFQIC AFG------- -RIQDEHDDQ ELIGIARQIA DKLKCSPLAA
LNAAH----- -VVSLDRLRS SDCWLLLKEV ALGG------ ---QPMDFPP ELQEILGAIV ANVKGLPLAT
VFCDQQQ--- -VVHLEKMDD ANFLALFKHH AFSGA----- -KIGDQLLHN KLEHTAVEIA KRLGQCPLAA
SSTSDEG--- FIYQMKPLSR NDSENLFLKR TLCAE----- ---DKFPV-- QLTGIKNDII EKCDGLPLAI
HHCGSSK-FV HNHKISLLSD NEAKELFLKK AFS------- ---SRNDYPQ HLEDVFAKVL RRCGGLPLAV
CSPPS----- HIHQPDFLGK EEARTLFLKK TNRS------ --LDELEKGD KTKGIVEKIL NKCGGLPLAI
ADS------- -VYKLQALSA GDSRKLFFLR IFGNE----- ---NRCLPK- ELDKESKNIL RKCGGVPLAI
IATMDPV--- ---HLDGLED DDFWLLFKSC VFG------- --DEKYEGHG NLQIIGQSIA KRLKGYPLAA
VKVKPTDS-- -TIELEGLDQ VAFREFFQSC VFGD------ -DNKSKDDHK ELDDIGEEIM KKLKGSPLAA
AVDH------ YRIELKTLQY AESWELFCKK AFVAL----- -KDSQCPE-- NLRFFAEKIV ARCQGLPLAL
VRST------ YQFCLPRLSS DDSWQLFQQS FRMP------ ---VKCLEP- GFIEVGKEIV ETCCGLPLAL
VTTIGCP--- --IRLERLSD EECMRFFQEC VFG------- -DQQTWEGHT NLHYYGCKIV KRLKGFPLAV
MGVDD----- -PMEVSCLQP EESWDLFQMK VGKNT----- --LGSHPD-- -IPGLARKVA RKCRGLPLAL
MGAEY----- -KLRVEFLEK KHAWELFCSK VWRKD----- --LLESSS-- -IRRLAEIIV SKCGGLPLAL
ADPT-----C LTFRASILNP EESWKLCERI VFPRR----- -DETEVRLDE EMEAMGKEMV THCGGLPLAV
HE--I----D LVYEVKLPSQ GLALKMISQY AFG------- ---KDSPPD- DFKELAFEVA ELVGSLPLGL
HG--I----N HVYKVEYPSN DEAFQIFCMN AFG------- ---QKQPYE- GFCDLAWEVK ALAGELPLGL
PYGIG----S TKHEIELLKE DEAWVLFSNK AFPGS----- --LEQCRTQ- NLEPIARKLL ERCQGLPLAI
ICCEQEH--- -VIHLKNMDD TEFLALFKHH AFSGA----- -EIKDQVLRT KLEDTAVEIA KRLGQCPLAA
VKCES----- DPHHLRLFRD DESWTLLQKE VFQGE----- ------SCPP ELEDVGFEIS KSCRGLPLSV
CSGKNGN--- -VHNLKVLKH NDALCLLSEK VFEEAT---- -YLDDQNNP- ELVKEAKQIL KKCDGLPLAI
ICCEQEQ--- -VIHLENMDD ADFLALFKHH AFSGA----- -KIGDQILCS RLEHTAEEIA KRLGQCPLAA
CASIAGSFHS LVYCLEPLQD HHAKELLLKK TNRS------ --HQALKIG- EAEHIFDMIL KKCAGLPLAL
TNYEH----- VVISLNRFDK IATTELFQQR VCKKES---N PEYNKDIEDG VRNKYQQDIF DTTQGLPLAL
IGAQD----- -VHRVELMSD DTGWELLWKS MNIN------ ----EEIEVA NLRGMGNEIV RMCGGLPLAL
AHPT------ RRLDIQPLGN AQAFDLFCRR TFYNE----- -KDHACPS-- DLVEVATSIV DRCQGLPLAL
N--------D IIYEVTALPD HESIQLFKQH AFG------- ---KEVPNE- NFEKLSLEVV NYAKGLPLAL
CCSSDGD--- SVYQMEPLSV DDSRMLFSKR IFPDE----- ---NGCIN-- EFEQVSRDIL KKCGGVPLAI
LNENQ----C KLYEVGSMSE QHSLELFSKH AFK------- ---KNTPPS- DYETLANDIV STTGGLPLTL
LNENQ----C KLYEVGSMSK PRSLELFSKH AFK------- ---KNTPPS- YYETLANDVV DTTAGLPLTL
MDSG------ -AIYMGILSS EDSWALFKRH SLEHK----- ---DP-KEHP EFEEVGKQIA DKCKGLPLAL
ASSGK----- PPHHMRLMNF DESWNLLHKK IFEKEG---- ------SYSP EFENIGKQIA LKCGGLPLAI
CCNYQWK--- FVYRMKPLDD YHSRQLFLRR IFGSG----- ---DRCPE-- PFEVLCEKIL QKCGGLPLAI
CCDYESD--- AIFKMETLGG NHSTELFFNR VFG------- ---FKHECSK QLKECSEEII RTCGGLPLAI
CCSNDYE--- YVYKMKALGT EDSRRLFFKR IFGSE----- ---DTCPS-- YLEEVSTGIL KRCGGLPLAI
ATGN------ CAIKLEPLGE KHSWKLFCKA AFRNS----- -DDKWCPS-- ELHDLATKFL QKCEGLPIAI
AKST------ RTILLKPLPE KEAWCLFCNT TFRED----- -ADRECPQ-- HLEHWALRIL NKCSGLPLAI
CCSDSYYG-- LVHEMKPLSE TDSERLLLAK AVGS------ --VDGCVPN- NIKLHCDEIL RRCDGIPLFI
ASSG------ CHLQLQPLGS SYALDLFCRR AFNNT----- -ADRKCPQ-- ELEDVAVSIV ERCKGLPLAI
CCK-DHG--- RIHYMKPLSD ADSRKLFFRR IFGTE----- ---DTCPP-- QFTEVSSEIL KKCGGLPLAI
AGG------- -AYNLEPLSM NNSRKLLYRR IFGTDSKDNN EDNEKCPD-- ELVEVSEKIL KKCAGVPLAI
IGG------- -AYSMKALCH ESSRKLFYTR IFG------N EEKYKCPDE- HLTEVSHRIL NKCAGVPLAI
VGS------- -PYNMKPLSQ NNSRKLLYKR IFGNEGKDNN EDIEKCPDA- ELTEVSERIL KKCAGVPLAI
SWW------- -CLQVESP-E DNSRKLLYRR VFGNENNNNV EDMGKCPIE- ELAEVSDRIL KKCAGVPLAI
CSSPFHD--- LVFKLRMLSE DDSKRLFFRR IFGSE----- ---DKCPH-- QLKDVSVEII KKCGGLPLAI
```

B-3

SORGHUM NBS SEQUENCE ALIGNMENT OF AMINO ACIDS COVERING GLPL AND RNBS-D MOTIFS FOR PHYLOGENETIC ANALYSIS

```
46          121
Apaf-1      CKGSPLVVSL IGALLRDFP- ------NRWE YYLKQLQ--- ---------- ----------
XA1         LKGNPLAAKT VGSLLGTNL- ----TIDHWD SIIKSEEW-- ---------- ----------
WS110C06B1  CSGVPLAIVT LAKMLATKM- ---GGKKEWH KVR------- ---------- ----------
TC90621     CAGSPLAATA VGSLLQTKT- ----SVDEWN AVLSKSAI-- ---------- ----------
TC89319     LKCSPLAAKT VGRLLIKKP- ----FQEHWM KILDNKEW-- ---------- ----------
TC86205     CSGSPLAIVT MASLLANKA- ---CTKQEWD RVC------- ---------- ----------
TC85900     LGQCPLAAKV LGSRLSTKK- ----DTAEWK GALKLR---- ---------- ----------
TC83499     CRGLPLALKV IGASLRDQP- -----PKIWL SAKNRLS--- ---------- ----------
TC81018     CGGQPLAIVT MAGLVACNQN KP---TKYWD KL-------- ---------- ----------
TC79065     CGGLPLAVVS IATKLAHKQS RD-----EWE KHG------- ---------- ----------
TC77858     CKCLPLALVT VGRAMSNKR- ----TPEEWS NALDTLKAS- ---------- ----------
TC76961     CGGLPLAILT IGAVLANKD- -----TEEWE NIY------- ---------- ----------
TC75876     CGGVPLAIIT ISSMLASKQ- ETENTSEYWA KVC------- ---------- ----------
RPS5        CRGLPLALNV IGEAMACKR- ----TVHEWC HAIDVLTS-- ---------- ----------
RPS2        CGGLPLALIT LGGAMAHRE- ----TEEEWI HASEVLTR-- ---------- ----------
RPP8        CGGLPLAVKV LGGLLANKH- ----TVPEWK RVS------- ---------- ----------
RPP5        VGSLPLGLSV LGSSLKGRD- -----KDEWV KMM------- ---------- ----------
RPP1        AGELPLGLKV LGSALRGMS- -----KPEWE RTL------- ---------- ----------
RPM1        CQGLPLAIAS LGSMMSTK-- ---KFESEWK KVY------- ---------- ----------
RP1D        LGQCPLAAKV LGSRLCRKK- ----DIAEWK AALKIG---- ---------- ----------
PRF         CRGLPLSVVL VAGVLKQKK- ---KTLDSWK VVEQSLS--- ---------- ----------
PIB         CDGLPLAIVV IGGFLANRP- ---KTPEEWR KLN------- ---------- ----------
NP853482    LGQCPLAAKV LGSRLSRKK- ----DIVEWK AALKLR---- ---------- ----------
N           AKGLPLALKV WGSLLHNLR- -----LTEWK SAI------- ---------- ----------
MLA6        CGGVPLAIIT IASALAGDQ- -KMKPKCEWD ILL------- ---------- ----------
M           TGGLPLTLKV TGSFLFRQE- -----IGVWE DTL------- ---------- ----------
L6          TAGLPLTLKV IGSLLFKQE- -----IAVWE DTL------- ---------- ----------
I2C-1       CKGLPLALKA LAGMLRSKS- ----EVDEWR NILRSEIW-- ---------- ----------
GPA2        CGGLPLAITL IAGLLSKIS- ---KTLDEWQ NVAENVR--- ---------- ----------
CD211851    CGGLPLALVS MAGYVGCNK- KPEELLKHFQ YLGPESTKDH REGLNQEEES TKDHRGGLNQ
BZ628476    CGGLPLAIIS IASILAIQPD ----NLELWR HVK------- ---------- ----------
BZ626449    CGGLPLAIIT LSSHLATQRD K--LDRELWE HTL------- ---------- ----------
BZ349832    CAGLPLVIVC ISSILATKG- ---KEATEWE KVY------- ---------- ----------
BZ349019    CVGSPLAAKA FGSMLSTKS- ----SIQEWK DMLAKSQI-- ---------- ----------
BZ342222    CEGLPIAIAC IGRLLSSKD- ---LTYAAWD SVY------- ---------- ----------
BZ329687    YGGVPLAIIT MASMLANKTG KEINAHNYWS NVY------- ---------- ----------
BM325057    CGGLPLALIT VGRAMASKR- ----TAKEWK HAITVLKI-- ---------- ----------
BM317647    CGGLPKVIVA VADFFAAG-- --------WR FNP------- ---------- ----------
BH246154    CSGLPLAIVS VGNVLALKE- ---KSEFAWK SVH------- ---------- ----------
BG557168    CKGLPLALKT MGGLMSSKH- ----QIKEWE AIAEDDRV-- ---------- ----------
BE596218    CKGLPLAIVT IGGFLAKQP- ---KTPIVWR KMN------- ---------- ----------
AAM94306    CAGVPLAIIT MASLLACKA- ---RNKMEWC KVC------- ---------- ----------
AAM94297    CAGVPLAIIT IASLLANKA- ---RDKMEWL EVY------- ---------- ----------
AAM94295    CAGVPLAIIT MASLLACKP- ---RNKMDWY EVC------- ---------- ----------
AAM94294    CAGVPLAIIT MASLLACKP- ---RNKMDWY EVY------- ---------- ----------
AAD27570    CGGLPLAIIT MASLLTTKS- ---DTRADWL KIC------- ---------- ----------
```

SORGHUM NBS SEQUENCE ALIGNMENT OF AMINO ACIDS COVERING GLPL AND RNBS-D MOTIFS FOR PHYLOGENETIC ANALYSIS (Continued)

```
-----NKQFK RIRKSSS--- -----YDYEA LDEAMSISVE MLRED-IKDY YTDLSILQKD V
---------- KSLQ------ ------QAYG IMQALKLSYD HLSNP-LQQC VSYCSLFPKG Y
-----GYIGS GLEN-TL--- ------DVKN MRMVTSLGYY NLPPH-LRAC LLYMSVFPED Y
---------- CDDE------ ------T--E ILPILKLSYN GLPPH-IRQC FAFCAIFPKD Y
---------- LEEN------ ------HDND IIPALKISYD YLPFH-LKKC FSCFCLFPDD Y
-----NSIGS TLEK-DP--- ------DVEE MRRILSLSFD DLPHH-LKTC LLYLSIFPED Y
---------- ---------- ------DLSE PFTVLLWSYK KLDPR-LQRC FLYCSLFPKG H
-------RGE AISDS----- ------HETK LLERMAASVE CLSEK-VRDC FLDLGCFPED K
----CKRLPA RETSVTEVFD KQVNSLTLEG VKRILDCCYN DLHGD-LKTC LLYLAMFPKG C
-------LNL LYNSRPE--- ------GLDG LKQILNLSYN DLQPQ-LKSC LLYLSIFPEN S
-------LPS GTPG------ ------LDKS THALVKFCYD NLESDMVREC FLTCALWPED H
-----MQLPW DLANNPS--- -------LDA LRRVVSLSYN HLPSR-LKPC FLHLSIFPED F
-----ESMGS GLENTSS--- ------DVID MRRILSVSYH DLPLH-LRTC LLYLSLYPED Y
-------SAI DFSG------ ------MEDE ILHVLKYSYD NLNGELMKSC FLYCSLFPED Y
-------FPA EMKG------ ------MNY- VFALLKFSYD NLESDLLRSC FLYCALFPEE H
-----DNIGS QIVGGSCLD- ----DNSLNS VYRILSLSYE DLPTH-LKHR FLFLAHFPEY S
---------P RLRNDS---- -------DDK IEETLRVGYD RLNKK-NREL FKCIACFFNG F
---------P RLRTSL---- -------DGK IGGIIQFSYD ALCDE-DKYL FLYIACLFNN E
---------S TLNWELN--- ---NNLELKI VRSILLLSFN DLPYP-LKRC FLYCSLFPVN Y
---------- ---------- ------DLSD PFTSLLWSYE KLDPR-LQRC FLYCSLFPKG H
-----SQRIG SLEES----- ---------- -ISIIGFSYK NLPHY-LKPC FLYFGGFLQG K
-----ENINA ELEMNP---- ------ELGM IRTVLEKSYD GLPYH-LKSC FLYLSIFPED Q
---------- ---------- ------DLSE PLTILLWSYK KLDPR-LQRC FMYCSLFPKG H
---------E HMKNNS---- -------YSG IIDKLKISYD GLEPK-QQEM FLDIACFLRG E
-----RSLGS GLTE-DN--- ------SLEE MRRILSFSYS NLPSH-LKTC LLYLCVYPED S
---------E QLRKTLD--- -------LDE VYDRLKISYD ALKAE-AKEI FLDIACFFIG R
---------E QLRRTLN--- -------LDE VYDRLKISYD ALNPE-AKEI FLDIACFFIG Q
---------- ELPS------ ------CSNG ILPALMLSYN DLPAH-LKQC FAYCAIYPKD Y
-----SVVST DLEAKC---- ---------- -MRVLALSYH HLPSH-LKPC FLYFAIFAED E
EEEPTKDHPG GLNQEEEPTK DHREGLNQEE AGRIISYCYN DMPAE-IKTC SLYLSIFPKG S
-----EALFS RLRYNLT--- ------SEVK LREIVSLSYN SLPCH-LKTC LLYLSMYPEG Y
-----NCLGS SLELNPT--- ------LEG- MRQILSMSYT NLPHC-LKAC VLYLGMYPED H
-----DSLGS GSND------ ------GLSW LWQAFEVSYD DLPQH-LKVC LLYLSAFRED Y
---------- CDER------ ------T--E IFPILKLSYD DLPSD-MKQC FAFCAVFPKD Y
---------R ELEFQP---- ---TNNVIRG VDIILKVSLE DLPYE-LKNC FLYCAIFPED Q
-----QAMGS GLDG-ST--- ------NVKN MRRILSVSYY DLPSH-LMNC FLYLSLFPED Y
-------APW QLLG------ ------MEFD VLEPLKKSYD NLPSDKLRLC LLYCSLFPEE F
-----IHLMK VLETNPAFT- ------CLRD MFAWVNSYFV SCPDS-LKPC IFYLSIFPVN H
---------D SLVWDES--- ---TDHGIGQ VSSILNLSID DLPYH-LKRC FLYCSIYPED F
---------- GKD------- ---------E VLSILKLSYM HLSSE-MKQC FAFCAVFPKD Y
-----EHISA ELEMNP---- ------ELGI IRAILMKSYD GLPYH-LKSC FLYMSIFPED Y
-----KSVGT GLEN-NS--- ------ALEN MRKILAFSYF DLPYH-LRTC LLYLSVFPED Y
-----NSIGT GLED-ST--- ------DVEN MRKILAYSYY DLKYH-LRVC LLYLSMFPED Y
-----NCIGT GLEN-SI--- ------DVEN MRKILSFSYY NMPSH-LRTC LLYFSVFPED Y
-----HSIGT GLQN-NL--- ------DVEN MRKILSFSYY NMPSH-LRTC LLYLSMFPED Y
-----NSIGC RLEK-NC--- ------DVEE MNMILSLSYN HLPHH-LKTC LLYLSMFPED Y
```

B-4

SORGHUM NBS SEQUENCE ALIGNMENT OF AMINO ACIDS COVERING RNBS-D AND MHDV MOTIFS FOR PHYLOGENETIC ANALYSIS

```
45          131
Apaf-1      DIKDYYTDLS ILQKDVKVPT KVLCILWDME TEEVEDILQE FVNKSLLFCD RNGKSFRYYL
XA1         PLQQCVSYCS LFPKGYSFSK AQLIQIWIAQ GFVE--E-SS EK------LE QKGWKY----
TC90621     HIRQCFAFCA IFPKDYEIDV EKLIQLWMAN GFIPEQHGVC PE------IT --GKKI----
TC89319     HLKKCFSCFC LFPDDYKFEK LEIICFWHSI GIIDYSR-QN KK------ME EIGSDY----
TC85900     RLQRCFLYCS LFPKGHKYKP DELVHLWVAE GLVGSCNLSS MT------IE DVGRDY----
TC83499     KVRDCFLDLG CFPEDKKIPL DVLINIWMEI HDLDEP--DA FA-----ILV ELSNKN----
TC81018     DLKTCLLYLA MFPKGCKTSR KCVTRRWIAE GFVTKK--YG LT------EE ELAETY----
TC80849     HLKECFLHCS LYPEEYPIQR FDLVRRWIAE GIVNPR---D NE-----LLE ESAEEY----
TC79065     QLKSCLLYLS IFPENSEIET KRLVRRWIAE GFIAGT--GS K--------E ETAISY----
TC76169     HLKTCLLYLS VFPEDYEIVK DRLIWRWIAE DFVPPGE-GG QS------SF ELGLSY----
RPS5        LMKSCFLYCS LFPEDYLIDK EGLVDYWISE GFINEK--EG RE-----RNI NQGYEI----
RPS2        LLRSCFLYCA LFPEEHSIEI EQLVEYWVGE GFLTSS--HG VN------TI YKGYFL----
RPP8        HLKHRFLFLA HFPEYSKISA YDLFNYWAVE GIYDG----- -S-----TIQ DSGEYY----
RPP5        KNRELFKCIA CFFNGFKVSN -----VKELL EDDVG----- ---------- ----------
RPP1        EDKYLFLYIA CLFNNE---S TTKVEEVLAN KFLDVG---- ---------- ----QG----
RPM1        PLKRCFLYCS LFPVNYRMKR KRLVRMWMAQ RFVEPIR--G VK------AE EVADSY----
RP1D        RLQRCFLYCS LFPKGHRYES NELVHLWVAE GFVGSCNLSR RT------LE EVGMDY----
PRF         YLKPCFLYFG GFLQGKDIHV SKMTKLWVAE GFVQAN--NE KG------QE DTAQGF----
PIB         HLKSCFLYLS IFPEDQIISR RRLVHRWAAE GYSTAA--HG KS------AI EIANGY----
NP853482    RLQRCFMYCS LFPKGHRYKP DELVHLWVAE GFVGSCISGR RT------LE DVGMDY----
N           KQQEMFLDIA CFLRGEE--K D-YILQILES CHIG------ ---------- --AEYG----
MLA6        HLKTCLLYLC VYPEDSMISR DKLIWKWVAE GFVHHEN-QG NS------LY LLGLNY----
M           EAKEIFLDIA CFFIGR--NK EMPYYMWSEC KFYP------ ---------- ---KSN----
L6          EAKEIFLDIA CFFIGQ--NK EEPYYMWTDC NFYP------ ---------- ---ASN----
I2C-1       HLKQCFAYCA IYPKDYQFRK EQVIHLWIAN GLVHQFHS-- ---------- --GNQY----
GPA2        HLKPCFLYFA IFAEDERIYV NKLVELWAVE GFLNEE--EG KS------IE EVAETC----
BZ626449    CLKACVLYLG MYPEDHEISK NDLVRQWVAQ GFISKA--GG QD------AE DIAVEY----
BZ423689    SLQQCFRYCC LFPKNYLFDA VKLVRMWISQ GFVHGNH-TG KK------LE DIGNAY----
BZ349832    HLKVCLLYLS AFREDYAIRR DRLTRRWITE GFVDEK--PG MS------MQ EVADNN----
BZ346314    QLKTCLLYLS IFPEDYQINK LRLIERWIAK GFVQQGD-GR QS------LH EIGQSY----
BZ342222    ELKNCFLYCA IFPEDQELTR RTLMRHWITS GFIKEKD--N RT------LE QVAEEY----
BZ338669    RLKPCFLYLS IFPEDYEIKR SHLVHRWIAE GFVRAK--VG TT------ID EVGKEY----
BZ337854    NLKTCLLYLS IFPEDYVIER ERLVRRWIAE GFICEE--RG LS------KQ EVAENN----
BM324406    RLQRCFAYCS IFPTTWRFNR YDLVKMWMAL GFIQPPTDEG KG------ME DLGQKY----
BM323307    EIITCSLYLG IFPKGSRISR KRLIRRWIAE GFVSEK--DG MS------VE DVAETY----
BH245455    HLKSCLLYLS VFPEDFSIDC RELILLWVAE GLIPGQ--DR ES------ME QLGRSY----
BG557168    EMKQCFAFCA VFPKDYEMDK DKLIQLWMAN NFIHADGTTD FV------QK --GEFI----
BE599136    QLQHCFSYCA LFPQDYKFEE VELINFWIGL NVLHSSHGDS KR------VE DIGESN----
BE596218    HLKSCFLYMS IFPEDYSISR RRLVHRWKAE GYSSEV--RG KS------KG EIADAY----
BE355823    NVKNCFLYCG LFPEDHQIRG EEIIRLWITE DFIEERGPTS IT------ME EVGAEY----
AAM94306    HLRTCLLYLS VFPEDYKISK NRLIWMWIAE GFIQSGR-HW GT------LF ACGESY----
AAM94297    HLRVCLLYLS MFPEDYPITK NHLIWMWIAE GFVQCE--QG KS------LF ELGECY----
AAM94295    HLRTCLLYFS VFPEDYKIEK HRLIWMWIAE GFIQCEK-HG ES------LF DLGESY----
AAM94294    HLRTCLLYLS MFPEDYEVEK DRLIWMWIAE GFIHCEK-QG KS------QY ELGENY----
AAD27570    HLKTCLLYLS MFPEDYVIKR DYLVRRWVAE GFISAH--GR KN------LE DEGECY----
```

SORGHUM NBS SEQUENCE ALIGNMENT OF AMINO ACIDS COVERING RNBS-D AND MHDV MOTIFS FOR PHYLOGENETIC ANALYSIS (Continued)

```
HDLQVDFLTE KNCSQLQDLH KKIITQFQRY HQPHTLSPDQ EDCMYWYNFL AYHMASAKMH KELCALMFSL D
------LAEL VNSGFLQQVE STRFS----- -S-------- ---------- ----EYFVMH DLMHDLAQKV S
------FMDL VSRSFFQDVN KVPFEVYDIE DP---R---- ---------- ----VTCKIH DLMHDLAQSS M
------LDEL VDSGFLIK-G DDN------- ---------- ---------- -----YYVMH DLLHDLSRTV S
------FNEM LSGSFFQLVS ETEYY----- ---------- ---------- ----SYYIMH DILHDLAQSL S
------LLTL VNDAQNKAGD LYSSYHD--- ---------- ---------- ----YSVTQH DVLRDLALHM S
------FNQL LRRKLIRPVD HSSNGK--L- ---------- ---------- ----KTFQVH DMVLDYIASK A
------YVEL ISRNLLQPDP ESVER----- ---------- ---------- -----CWITH HLLRSLARLL I
------LNEL IGRNLVQPLD LNHDNIP--- ---------- ---------- ----RRCTVH PVIYDFIVCK S
------FNDL VNRSLIQPAD MD-DEG-TP- ---------- ---------- ----ISCRVH DMVLDLICNI S
------IGTL VRACLLLEEE RN--KS---- ---------- ---------- -----NVKMH DVVREMALWI S
------IGDL KAACLLETGD E---KT---- ---------- ---------- -----QVKMH NVVRSFALWM A
------LEEL VRRNLVIADN RYLSSHS--- ---------- ---------- ----KNCQMH DMMREVCLSK A
------LTML AEESLIRITP VG-------- ---------- ---------- -----YIEMH NLLEKLGREI D
------IHVL AQKSLISFEG E--------- ---------- ---------- -----EIQMH TLLEQFGRET S
------LNEL VYRNMLQVIL WNPFGR--P- ---------- ---------- ----KAFKMH DVIWEIALSV S
------FNDM VSVSFFQLVF HIYCD----- ---------- ---------- ----SYYVMH DILHDFAESL S
------LDDL IGRNVVMAME KRPNTK---- -----V---- ---------- ----KTCRIH DLLHKFCMEK A
------FMEL KNRSMILPFQ QSGSSRKSI- ---------- ---------- ----DSCKVH DLMRDIAISK S
------FNDM VSGSLFQMVS QRYFV----- ---------- ---------- ----PYYIMH DILHDLAESL S
------LRIL IDKSLVFISE YN-------- ---------- ---------- -----QVQMH DLIQDMGKYI V
------FNQL INRSMIQPIY NYSGEA---- ---------- ---------- ----YACRVH DMVLDLICNL S
------IIFL IQRCMIQVGD DG-------- ---------- ---------- -----VLEMH DQLRDMGREI V
------IIFL IQRCMIQVGD DD-------- ---------- ---------- -----EFKMH DQLRDMGREI V
------FIEL RSRSLFEMAS EP--SERDVE EF-------- ---------- ----L---MH DLVNDLAQIA S
------INEL VDRSLISIHN VSFDGET--- ---------- ---------- ----QRCGMH DVTRELCLRE A
------FNEI VNRSIIQPAH TDSNN--DV- ---------- ---------- ----LSCRVH DMMLDLIIHK C
------LADL VNSGFLVNLG FIKLVGRGRN YS-------- ---------- ----NHFVMH DLMHDLAWEV S
------FTEL IGRNMIQAVD --VDCFGEI- ---------- ---------- ----HACKIH DVMFDLITKK S
------FNEL LNRSLIQPAD LDEDEM-NL- ---------- ---------- ----FSCRVH DMVLDLICSL S
------LNDL VNRSLLQVVI KNASGR--V- ---------- ---------- ----KRCRMH DVIRHLAIEK A
------FDEL ISRSMIQSSE --LGMEGSV- ---------- ---------- ----KTCRVH DIMRDIIVSI S
------FYEL INKSMVQPVD --VGYDGKA- ---------- ---------- ----RACQVH DMMLELIISK S
------FDDL LSRSFFGTAN KDQQ------ ---------- ---------- ----TYYFLD DLMHSLAQHF S
------FGHL VRRKMIRPVE HSSSGR--I- ---------- ---------- ----KQCVVH DMVLEHIVSK A
------LNEL INRSLVQPTK VGVDGT-NV- ---------- ---------- ----KQCRVH DVILEFIVSK A
------FSEL VWRSFIQDVD VKIFDEYHFA APAHKK---- ---------- ----IGCKMH DLMHDLAQET T
------LREL VNHGFLEKEG EKDGK----- ---------- ---------- ----SCYIIH DLLHDLARKV S
------FMEL IERSMVLPSK ESIGSRKGI- ---------- ---------- ----SSCKLH DLMREISISK A
------LNEI AQRSLLQVVQ RDAYGR--S- ---------- ---------- ----EIFQMH DLVRDIVVSK S
------FNEL INRSMIQPIH DTDTG--LI- ---------- ---------- ----KQCRVH DMILDLICSL S
------FNEL INTSMIQPVY DRHEA--MI- ---------- ---------- ----EHCRVH DMVLEVIRSL S
------FNEL ISRSMIQPIH GYNND--TI- ---------- ---------- ----YECRVH DMVLDLICSL S
------FNEL INRSMIQPIY GVSSN---V- ---------- ---------- ----YECRVH DMVLDLICSL S
------FNEL INRSLIQPVD FQYDGR--V- ---------- ---------- ----YTCRVH DVILDLITCK A
```

VITA

NAME: Jae-Min Cho

DATE OF BIRTH: February 8, 1967

PERMANENT ADDRESS: Texas A&M University
Dept. of Plant Pathology & Microbiology
2132 TAMU
College Station, TX 77843-2132

EDUCATIONAL BACKGROUND:

DEGREE: Master of Science
MAJOR SUBJECT: Plant Pathology
UNIVERSITY: Kyungpook National University (Korea)
DATE OF GRADUATION: February 1997

DEGREE: Bachelor of Science
MAJOR SUBJECT: Agricultural Biology
UNIVERSITY: Kyungpook National University (Korea)
DATE OF GRADUATION: February 1993

PROFESSIONAL BACKGROUND:

INSTITUTE: National Yeongnam Agricultural Experiment Station (Korea)
FINAL POSITION: Agricultural Researcher
DURATION: July 1992 – June 1999
JOB DESCRIPTION: Development of Plant Protection Strategy